Statistica e ricerca a di mercato - Statistica
Prof. Zavarrone

16/02

Introduzione alla statistica
Def. **statistica**= la statistica è la scienza che permette di raccogliere, organizzare, sintetizzare e analizzare le informazioni al fine di pervenire ad una o più conclusioni. In altri termini, la statistica fornisce risposte a specifiche domande di ricerca. La statistica *fornisce, inoltre, una misura di probabilità associata ad ogni conclusione raggiunta.*

Statistica= data storytelling:
- i numeri raccontano una storia. Non sono semplicemente strumenti di calcolo, ma mezzi per dare forma e forza a un racconto. Ogni informazione quantitativa aggiunge una dimensione oggettiva a ciò che descriviamo. (Es. se parliamo di un brand che detiene il 33% della quota di mercato, non stiamo solo citando un numero: stiamo raccontando che quel brand occupa un terzo del mercato, lasciando gli altri due terzi ai concorrenti, che possono essere frammentati, più piccoli o persino più grandi.
- Il numero diventa un elemento narrativo che aiuta a comprendere la struttura della realtà.

Differenza tra matematica e statistica
In questo senso si chiarisce anche la differenza tra matematica e statistica:
- La matematica è ciò che ci permette di fare calcoli; la statistica, invece, ci insegna a ragionare sui dati, a interpretarli, a inserirli in un contesto e a utilizzarli per prendere decisioni consapevoli. L'obiettivo non è semplicemente ottenere un risultato numerico, ma sviluppare un pensiero critico capace di comprendere cosa quel numero significhi davvero. Studiare statistica significa imparare a decidere in modo informato, valutando le implicazioni delle proprie scelte.

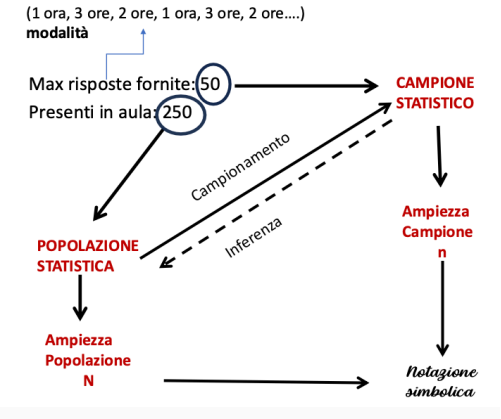
Statistica ai tempi di catgut & gemini
Domanda: Perchè devo studiare statistica se vado su Chatgpt o Gemini e ho tutte le risposte?
Risposta: Le informazioni fornite sono suggerimenti o dettagli operative e per poterli adeguatamente usare si necessita di INTERPRETAZIONI che possano mettere in relazione il contesto con i dati

La statistica fornisce proprio questo: gli strumenti per capire cosa sta succedendo e per valutare criticamente le informazioni.

Metafora del pentagramma: il dato è come una nota musicale: da solo ha un suono, ma è solo quando viene collocato su un pentagramma, in relazione ad altre note, che genera una melodia (il contesto è il pentagramma; i dati sono le note).
=> la stessa informazione, inserita in contesti diversi, può produrre interpretazioni differenti. Ciò che conta non è solo l'elaborazione tecnica del dato, ma la sua interpretazione, che nasce dall'unione tra metodo, contesto, sensibilità e creatività.

Variabili e modalità
Una variabile è l'aspetto che vogliamo rilevare; le modalità sono le sue possibili manifestazioni. Se la variabile è il colore dei capelli, le modalità possono essere biondo, castano, nero, rosso. Se la variabile è il reddito, le modalità saranno valori numerici differenti. Le variabili possono essere qualitative o quantitative, ma in entrambi i casi è essenziale individuarle con precisione.

Statistica: popolazione e campione
La popolazione statistica è l'insieme delle unità esposte allo stesso fenomeno nello stesso momento. Ad esempio, tutti gli studenti presenti in aula costituiscono una popolazione statistica se vogliamo misurare una loro caratteristica comune. Ogni individuo è un'unità statistica. Spesso, però, non è possibile analizzare l'intera popolazione, quindi si seleziona un campione, cioè una parte della popolazione. Affinché i risultati siano validi e generalizzabili, il campione deve essere scelto con criteri rigorosi e non "a casaccio". La selezione deve essere casuale nel senso statistico del termine, cioè guidata da regole e probabilità, non da preferenze personali o autoselezione.



Notazione simbolica
Ampiezza Campione → n
Ampiezza della popolazione → N

Come dalla popolazione possiamo estrarre un campione statistica, il risultato che otteniamo dal campione statistico ci consente di ricavare la popolazione statistica.

Cultura del dato

Caso Coca-Cola degli anni '80, che mostra quanto sia pericoloso ignorare il contesto.

L'azienda, per contrastare la concorrenza di Pepsi, decise di modificare la formula della Coca-Cola tradizionale sulla base di test di assaggio che sembravano indicare una preferenza per un gusto più dolce. Dal punto di vista metodologico, l'indagine era stata condotta correttamente. Tuttavia, fu trascurata una dimensione fondamentale: il significato simbolico della Coca-Cola per il consumatore americano. La bevanda non era solo un prodotto, ma un simbolo identitario e culturale. Ignorando questo aspetto, l'azienda sottovalutò la reazione emotiva e sociale del pubblico. Il risultato fu un fallimento e un forte danno reputazionale.

17/02

Caso 1 - Aerei: I dati rilevano informazioni
Durante la II guerra mondiale, gli arei tornavano dalla battaglia pieni di buchi di proiettile sulle ali e sulla coda. I generali volevano rinforzare proprio quelle zone. Un matematico, statistico Abraham Wald disse: "Assolutamente no, dovete rinforzare i motori, dove non c'è nemmeno un buco."
Questo perché dall'osservazione di quello che succedeva agli aerei, i dati hanno evidenziato che i danni derivavano dal fatto che gli aerei erano lenti. Una maggiore velocità dei mezzi avrebbe evitato il proiettile.
Questo è il caso in cui i **dati rilevano informazioni.**

Nella lezione precedente abbiamo visto 3 casi negativi dove informazioni di tipo statistico ha creato dei danni. In questo caso vediamo la parte pro-ATTIVA.

Caso 2 - Cioccolato: Statistica e fake news
Per anni si è alimentata la credenza secondo cui mangiare cioccolato fa dimagrire, successivamente si è passati a consigliare cioccolato fondente puro senza zuccheri.
Questa affermazione deriva da una fake news creata con dati statistici.
Questo accade perché **con la statistica si creano le fake news** ma se usata bene può smentirle.

La strategia del "giornalista scientifico"
Il caso riguarda un giornalista con un passato da ricercatore che ha voluto dimostrare quanto fosse facile creare una notizia falsa con l'apparenza di una ricerca scientifica seria. Per farlo, ha condotto uno studio "estremizzato" con l'obiettivo di sostenere che il cioccolato facesse dimagrire.
Come ha ottenuto il risultato desiderato (le "Leve")
Per mascherare la falsità dei dati e ottenere un risultato che sembrasse statisticamente rilevante, ha utilizzato due leve principali:

1. **Campioni molto piccoli:** Invece di testare il cioccolato su una popolazione ampia, ha usato pochissime unità statistiche. I campioni ridotti sono più volatili e soggetti a variazioni casuali che possono apparire come "scoperte".

2. **Un numero enorme di variabili:** Sulle poche persone testate, sono stati misurati tantissimi aspetti (le **variabili**[1]), come peso, colesterolo, umore, pressione, ecc..

La tecnica del p-hacking e la "tortura dei dati"
Il giornalista ha ripetuto l'esperimento più volte, cambiando continuamente metodologie analitiche. In statistica, se si misurano troppe variabili su un campione troppo piccolo, aumenta drasticamente la probabilità di trovare una correlazione puramente casuale che sembri significativa.
Questo comportamento scorretto, definito **p-hacking,** consiste nel manipolare o 'torturare' i dati ripetendo l'esperimento e cambiando metodologie finché non emerge il risultato voluto.
Come afferma una citazione celebre, si tratta di: **"Tortura i dati finché vuoi e qualcosa ti diranno"**.
In pratica, se si elaborano i dati abbastanza a lungo e con metodi diversi (non corretti), alla fine i numeri "confesseranno" quello che il ricercatore vuole sentirsi dire, anche se il risultato è scientificamente nullo.

Perché il risultato era nullo?
Secondo i pilastri della statistica, se il **numero di variabili è superiore all'ampiezza campionaria** (cioè ci sono più cose misurate che persone nel test), la prassi metodologica è scorretta e il significato del risultato è pari a zero. Nonostante ciò, lo studio fu pubblicato su una rivista scientifica, alimentando per anni la falsa credenza che il cioccolato aiuti a perdere peso.
In sintesi, la stessa statistica che può essere usata per creare queste **fake news** è lo strumento necessario per **smascherarle**, a patto che venga applicata con rigore scientifico e campioni adeguati.

Campione e campionamento

In questa lezione parleremo di campioni ed è fondamentale chiedersi qual'è la validità di essi, come si devono costruire.

La statistica interviene quando vogliamo conoscere e quantificare un fenomeno. Il suo obiettivo è proprio quello di cercare un "fil rouge" (un filo conduttore) tra le diverse manifestazioni di un fenomeno per poterle sintetizzare. Poiché non è possibile presentare ogni singola manifestazione per ogni unità, la statistica offre una sintesi che fornisce un'idea e un'interpretazione del fenomeno stesso.
In questo modo ci fornisce delle interpretazioni dandoci un'idea.

Per arrivare a questa sintesi, si può lavorare sulle unità statistiche in due modi:

- In modo estensivo → **Indagini totali (o censuarie)**: consistono nel fissare la popolazione di interesse e misurare le variabili su ogni singola unità statistica che la compone. *Es. il censimento, una pratica che accompagna l'umanità fin dai tempi antichi (come il censimento citato per la nascita di Gesù*).

  Si ricorre alle indagini totali anche quando si studiano le nicchie di mercato, ovvero una parte ristretta di unità statistiche accomunate dagli stessi comportamenti.
  Questo perché, se la popolazione è già piccola, estrarre un campione porterebbe ad avere un numero di unità troppo esiguo (scarso, trascurabile).
  In questo caso, si ricadrebbe nell'errore del p-hacking: un campione troppo piccolo su una popolazione limitata produce risultati spesso errati o opposti alla realtà.
  Pertanto, per popolazioni di dimensioni contenute, è sempre consigliabile l'indagine totale.

- Per porzioni → **Indagini campionarie:** si lavora solo su una porzione della popolazione.

[1] **variabili**: diversi aspetti che vogliamo rilevare sulle unita statistiche di riferimento

Le indagini campionarie sono necessarie quando la popolazione è molto ampia e bisogna selezionare delle unità attraverso un piano di campionamento.

Il **piano di campionamento*** riassume due grandi categorie di approcci:
- **Campionamento probabilistico** (applicazione di criteri = equirappresentatività di ogni unita inserita nel campione)
- **Campionamento non probabilistico** (soggettivata del ricercatore)

Gli elenchi (o database) utilizzati per le indagini statistiche e le ricerche di mercato vengono acquistati e costruiti principalmente sulla base dei dati che gli utenti lasciano volontariamente durante la navigazione o l'iscrizione a servizi online.
Come si alimenta questo mercato:

- Cessione dei dati a terzi: Il meccanismo principale scatta quando, iscrivendosi a un sito, l'utente mette la spunta sull'autorizzazione a ricevere informazioni o a cedere i propri dati a terzi. Quella che per l'utente è solo una "spunta" veloce, per le società di dati rappresenta un valore economico di migliaia di euro, poiché ogni contatto ha un prezzo elevato sul mercato.
- Aggregatori e Panel: Gli elenchi vengono acquistati da aggregatori di punteggio o attraverso panel precostituiti (come quelli citati di *SurveyMonkey*), che sono pacchetti di rispondenti già pronti per essere veicolati in diverse indagini.
- Fonti in evoluzione: Mentre in passato si faceva ampio ricorso agli elenchi telefonici, oggi i database vengono alimentati costantemente dalle tracce lasciate digitalmente.

Questo mercato di vendita di dati va ad alimentare i database utilizzati poi nei campionamenti non probabilistici (come quello a quote o di convenienza), per questo è importante non regalare informazioni personali con leggerezza.
Nelle fonti viene spiegato che l'algoritmo di Facebook ai suoi inizi (anche se oggi non funziona più così) operava proprio secondo la logica del campionamento snowball o a valanga.
Ecco come veniva applicato questo concetto al social network:

- All'inizio, l'algoritmo chiedeva all'utente di scorrere la propria rubrica e poneva domande del tipo: "Conosci questo? Conosci quest'altro?".
- Una volta che l'utente indicava di conoscere una persona, quel contatto veniva importato e l'algoritmo andava a "curiosare" nei contatti di quella stessa persona per proporre altri nominativi.
- La crescita a valanga = Proprio come una palla di neve che rotolando diventa sempre più grande, il sistema creava una catena di persone che si conoscevano tra loro, unite dallo stesso interesse o contesto sociale.

In statistica, questo metodo viene usato per studiare **popolazioni "sfuggenti"**[2] (come gli immigrati clandestini o chi esercita la prostituzione), dove non esiste un registro ufficiale. Si parte da un singolo individuo che fornisce altri nominativi, i quali a loro volta ne forniscono altri ancora, permettendo di quantificare realtà che altrimenti rimarrebbero invisibili.
Sebbene sia un metodo non probabilistico e quindi i risultati non siano generalizzabili a tutta la popolazione, è uno strumento fondamentale quando mancano elenchi completi da cui estrarre i dati.

Concetti chiave generali: perché ricorrere al campionamento?
Oltre a motivi di tempo e di costo legati a popolazioni molto ampie, il campionamento è indispensabile quando l'osservazione del fenomeno implica la **distruzione delle unità statistiche**.
Tre esempi pratici:

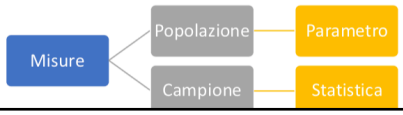
- **Lampadine:** per testare la durata media di un lotto, bisognerebbe tenerle accese finché non si bruciano; se si testasse l'intera produzione, l'azienda non avrebbe più nulla da vendere.

[2] **Popolazioni "sfuggenti":** Esistono gruppi di persone (chi esercita la prostituzione, gli immigrati clandestini o i lavoratori in nero) che non sono iscritti in nessun registro ufficiale o anagrafe. Poiché non esiste una lista da cui estrarre i nomi casualmente, lo statistico non può usare i metodi classici

- **Farmaci:** i test clinici non possono essere eseguiti su tutta la popolazione.
- **Qualità del cibo:** per valutare un alimento, spesso è necessario cucinarlo o consumarlo, rendendone impossibile la vendita.

Parametri vs Statistiche (L'analogia del cappello)

DOMANDA ESAME - Come capire se un'indagine è totale o campionaria dal nome delle misure.



Parametro	Statistica
è una misura calcolata sulla **popolazione** (es. la media della popolazione).	è una misura calcolata su un **campione**.
L'analogia del *colbacco* "Quando fa freddo" (Inverno) → Colbacco → Popolazione → Parametro L'idea è che d'inverno ci si metta un cappello pesante e ingombrante come il colbacco. Questo cappello "grande" rappresenta la popolazione (l'insieme totale). Le misure calcolate su tutta la popolazione si chiamano ufficialmente parametri.	L'analogia del *berretto estivo* "Quando fa caldo" (Estate) → Berretto → Campione → Statistica D'estate si usa un berretto leggero, che rappresenta il campione (una piccola parte del tutto). Le misure calcolate su un campione si chiamano ufficialmente statistiche.
Se in un testo leggi: "Il parametro della media è 28", sai subito che si tratta di un'indagine totale (censimento) perché è stato usato il termine "parametro".	Se leggi: "La statistica media è 35", sai che si tratta di un'indagine campionaria perché è stato usato il termine "statistica"

Rappresentatività e dimensione del campione

Un campione è valido solo se è una "**miniatura**" della popolazione, ovvero se ne riflette fedelmente le caratteristiche.
Viene usato l'*esempio del Cicaplast*: il campioncino gratuito è piccolo, ma il prodotto all'interno deve essere identico a quello della confezione standard.

Le tre leve per determinare l'ampiezza del campione

Per stabilire quanto deve essere grande un campione non esiste un numero fisso, ma bisogna considerare tre fattori fondamentali:

- **Dimensione della popolazione di partenza:** Quante sono le unità totali.
- **Margine di errore:** Quanto rischio di sbagliare si è disposti ad accettare per avere un campione.
- **Livello di probabilità:** La verosimiglianza dei risultati. Nelle indagini socio-economiche si usa solitamente il **95%**; non si usa il 100% perché equivarrebbe a intervistare tutta la popolazione.

Dobbiamo nominare queste 3 leve per avere informazioni sulla dimensione campionaria.

Alcuni studi statistici indicano il **384** come il **numero perfetto** per un campione, indipendentemente dalla grandezza della popolazione di partenza.
Questa affermazione risulta "un po' forte". Pur essendo un valore di riferimento spesso usato, sia è meglio utilizzare dei **simulatori o app** (come quello di SurveyMonkey) che calcolano la dimensione esatta incrociando i tre fattori sopra citati.

***Differenza tra piani di campionamenti:**

- **Il campionamento probabilistico:** si fonda sul criterio dell'**equiprobabilità** e dell'**equirappresentatività**. Ciò significa che ogni unità della popolazione ha la stessa probabilità di essere estratta per far parte del campione. In questo approccio non interviene la

scelta soggettiva del ricercatore, ma si applicano criteri matematici che garantiscono che il campione sia una miniatura fedele della popolazione.
I metodi di campionamento probabilistico:

1. **Casuale semplice**: Ogni elemento della popolazione ha la stessa probabilità di essere campionato.
2. **Stratificato**: si suddivide la popolazione in strati omogenei all'interno e si seleziona casualmente un certo numero di elementi da ogni strato.
3. **Sistematico**: Si divide la popolazione in k gruppi (N/n) e si estrae casualmente un primo elemento dal 1° gruppo ed i restanti vengono estratti ogni k dal primo
4. **A grappolo**: si suddivide la popolazione in gruppi eterogenei all'interno e si seleziona casualmente un grappolo

- **Campionamento non probabilistico**: in questo approccio, la selezione delle unità è guidata dalla soggettività del ricercatore. I risultati ottenuti non possono essere generalizzati all'intera popolazione e sono validi solo per il gruppo specifico e il momento in cui sono stati rilevati.
  Le principali tipologie sono:
  1. **Campione di convenienza** si scelgono le unità più facilmente raggiungibili (es. gli studenti presenti in aula in quel momento) perché non richiede sforzo o costi elevati
  2. **Campione a scelta ragionata** il ricercatore seleziona i rispondenti in base a specifiche competenze o caratteristiche necessarie per lo studio (es. un team di esperti o autisti di una specifica tratta negli orari di punta)
  3. **Campione a quote*** è il metodo più simile a quelli probabilistici ed è molto usato nelle ricerche di mercato. Il campione deve rispecchiare le stesse proporzioni (quote) della popolazione (es. se nella popolazione il 25% sono giovani, il campione dovrà avere il 25% di giovani). Tuttavia, una volta fissata la quota, la selezione delle singole persone non avviene con metodo scientifico, ma fino a esaurimento dei posti disponibili
  4. **Campione snowball o effetto a valanga** si parte da un piccolo gruppo di persone che suggeriscono altri nominativi, creando una catena.

*__Esempio campionamento a quote__
Il campionamento a quote è un metodo non probabilistico molto utilizzato nelle ricerche di mercato perché permette di costruire un campione che sia una "miniatura" della popolazione, pur senza utilizzare criteri di estrazione casuale.

1. Definizione delle variabili e delle proporzioni
   Immaginiamo di voler testare una nuova crema per il viso. Il ricercatore identifica prima di tutto le caratteristiche rilevanti della popolazione, come le fasce d'età. Ad esempio, si osserva che nella popolazione di riferimento:
   - Il 25% è composto da donne giovani (18-25 anni).
   - Il 75% è composto da donne mature (oltre i 25 anni).

2. Costruzione del campione
   Se decidiamo di intervistare 100 persone, il metodo a quote impone che il campione rifletta esattamente le percentuali della popolazione reale (le cosiddette "quote"). Quindi dovremo selezionare:
   - 25 donne giovani.
   - 75 donne mature.

3. La fase di rilevazione (Saturazione delle quote)
   I selezionatori iniziano a contattare le persone da elenchi o database (spesso derivanti da dati ceduti online o ricerche di mercato precedenti). Il meccanismo funziona così: se una donna matura risponde alla chiamata, le viene chiesta l'età. Se rientra nella fascia desiderata e la quota dei 75 posti non è ancora stata raggiunta, viene intervistata.
   Tuttavia, nel momento in cui viene raggiunta la quota (ad esempio, abbiamo già 75 interviste di donne mature), la quota si dice "chiusa" o saturata. Se dovesse rispondere una 76ª donna matura, questa verrebbe scartata perché non c'è più spazio per quella specifica categoria nel campione.

4. Perché non è un metodo probabilistico?

Nonostante il campione sembri molto preciso perché rispetta le proporzioni della realtà, è considerato non probabilistico per due motivi principali:

- Mancanza di casualità: Non viene usato un metodo scientifico (come l'estrazione casuale) per scegliere chi chiamare; si scorre semplicemente una lista finché non si riempiono i posti disponibili.
- Soggettività e convenienza: La probabilità di essere estratti non è uguale per tutti, poiché dipende da chi è presente negli elenchi o da chi risponde per primo al telefono.

In sintesi, il campionamento a quote cerca di salvare il "rapporto percentuale" della popolazione, ma perde il rigore scientifico dell'equirappresentatività tipico dei metodi probabilistici.

CAMPIONAMENTO PROBABILISTICO

Quando ciascuna unità della popolazione è selezionata in **maniera casuale** e ha una **probabilità positiva e nota** di entrare a far parte del campione.

Il campionamento probabilistico si basa su due principi fondamentali: **equiprobabilità** ed **equirappresentatività**.

Significa che ogni unità della popolazione ha la stessa probabilità (nota e diversa da zero) di essere scelta per entrare nel campione. In questo metodo non interviene la scelta soggettiva del ricercatore: la selezione avviene attraverso criteri matematici e procedure casuali, così che il campione sia una vera miniatura fedele della popolazione.

Grazie a queste caratteristiche, i risultati possono essere generalizzati all'intera popolazione con maggiore affidabilità.

Il concetto matematico alla base è che ciascuna entità della popolazione viene selezionata in modo casuale e ha quindi una probabilità positiva (+ 0) e nota, di entrare a far parte del campione.

Queste tecniche probabilistiche sono dei pilastri dell'analisi campionaria poiché garantiscono che il campione che viene costruito (sempre se la tecnica è correttamente applicata) è rappresentativo della popolazione di riferimento.

I metodi di campionamento probabilistico:

1. **Campionamento causale semplice:** È il metodo basico e più importante. Ogni elemento della popolazione ha la stessa probabilità di essere selezionato.

   Analogia della tombola:
   - La popolazione: È rappresentata dai 90 numeri nel sacchetto (il tabellone).
   - Il campione: È rappresentato dalle cartelle, che sono espressioni in piccolo della popolazione totale.
   - Il processo: Si assegna un codice a ogni unità della popolazione (es. un elenco di 1000 persone) e si utilizza un generatore di numeri casuali per estrarre le unità che comporranno il campione.

   Esistono due modalità di estrazione:
   - Con reinserimento: Estraendo più campioni posso trovare la stessa unità (es. persona) in più campioni differenti.
   - Senza reinserimento: Ogni unità (es. una persona) può comparire una sola volta quindi ogni campione sarà composto da unità differenti. Questo è il metodo tipicamente usato nelle scienze economiche e sociali.

   Se abbiamo una popolazione di 1000 persone e vogliamo costruire un campione di 10, ci possono essere più campioni.

   Per calcolarle si usa un principio del calcolo combinatorio, cioè il **coefficiente binomiale** (N/n*). La formula:
   - N = ampiezza della popolazione
   - n = ampiezza del campione

   $\binom{N}{n}$

   *Lettura: si legge "N su n piccolo".

Il calcolo si basa sul concetto di **fattoriale**, indicato dal **punto esclamativo (!)**. Il fattoriale è un "artificio" matematico che consiste nel moltiplicare un numero naturale per tutti i numeri precedenti fino ad arrivare a 1.
Esempio di fattoriale: 9! (9 fattoriale) significa fare 9×8×7×6×5×4×3×2×1.

Esempio di coefficiente binomiale: una popolazione (N) composta da 9 lettere da cui vogliamo estrarre campioni (n) di ampiezza 3. La formula applicata è:

1. Calcolo N! (9 fattoriale) quindi vado indietro moltiplicando 9×8×7×6×5×4×3×2×1.
2. Al denominatore abbiamo n! = 3! (3 fattoriale) quindi 3x2x1, tutto moltiplicato per (N - n)! quindi (9-3)! quindi 6! (6 fattoriale) quindi 6x5x4x3x2x1.
3. Semplificando otteniamo 84
4. Quindi esistono 84 modi diversi di combinare 3 lettere partendo da un gruppo di 9.

N=a,b,c,d,e,f,g,h,j
n_1= a,b,c
n_2= a,b,d
n_3= a,b,e
n_4= a,b,f
.....

$$\binom{N}{n} = \frac{N!}{n!\,(N-n)!}$$

$$\binom{9}{3} = \frac{9!}{3!\,6!} = \frac{\cancel{9}^{3} * \cancel{8}^{4} * 7 * \cancel{6} * 5 * \cancel{4} * \cancel{3} * \cancel{2} * \cancel{1}}{\cancel{3} * \cancel{2} * \cancel{1} * \cancel{6} * \cancel{5} * \cancel{4} * \cancel{3} * \cancel{2} * \cancel{1}} = 84$$

Perché calcolare tutti i campioni possibili?
Questo serve per l'inferenza statistica:

- Gestione dell'errore: Ogni singolo campione può produrre una misura che sovrastima o sottostima il dato reale della popolazione.
- Correzione delle stime: Sapere quanti campioni sono estraibili permette di capire che la media delle medie di tutti i campioni possibili corrisponde esattamente alla media della popolazione.
- Pensiero critico: Comprendere questo calcolo serve a capire che per fare inferenza non basta "un campione che funziona", ma bisogna lavorare in termini di probabilità e correzione dell'errore.

2. Campionamento sistematico[3]

Si divide la popolazione in k gruppi (N/n) e si estrae casualmente un primo elemento dal 1° gruppo ed i restanti vengono estratti ogni k dal primo. Questo significa che ogni unità della popolazione ha una probabilità nota di essere selezionata.

Può essere utilizzato anche quando non abbiamo una lista perfettamente strutturata, purché conosciamo due informazioni fondamentali:

- N = grandezza della popolazione
- n = ampiezza del campione che vogliamo estrarre

Come funziona?
Si mette a confronto N con n facendo il rapporto: $k = \frac{N}{n}$

Il valore **k** è chiamato **passo di campionamento** e indica ogni quante unità dobbiamo selezionare un individuo. Il termine "sistematico" significa proprio questo: si segue una regolarità fissa. Prima però si deve scegliere un numero iniziale casuale, chiamato **p**, compreso tra 1 e k. Questo numero rappresenta il **punto di partenza**. Poi si procede aggiungendo ogni volta il valore k.

Formula = ***p, p + k, p + 2k, ….. p + (n-1)k***

Esempio 1:
N: 150 persone in aula
n: 15 ampiezza campione

[3] Fase 1: approssimare l'ampiezza della popolazione
Fase 2: determinare l'ampiezza desiderata del campione n
Fase 3: calcolare N/n e arrotondare all'intero più vicino
Fase 4: selezionare a caso un numero compreso tra 1 e k. Chiamare questo numero p
Fase 5: il campione sarà composto dai seguenti individui.

k: 10 passo di campionamento (150/15)
Compreso il passo di campionamento bisogna estrarre 1 persona ogni 10. Si sceglie casualmente un numero tra 1 e 10 (questo numero prende il nome di “p”).

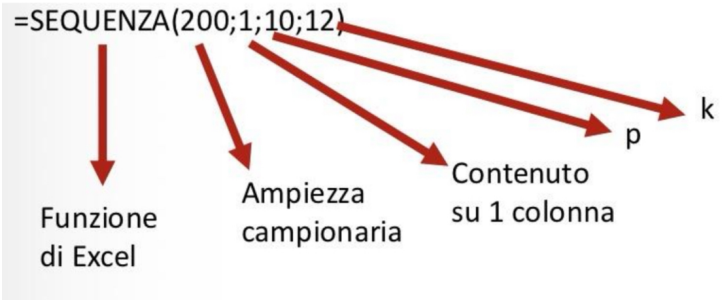
Esempio 2:
N: 248 n: 40
N/n: 248/40 = 6,2 approssimato a 6
Scegliamo come p il numero 5.
5, 5 + 6, 5 + 12, …, 5 + (40 – 1) * 6
5, 11, 17, …, 239 => 239 è la posizione dell’ultima unita che devo includere nel mio campione.

Esempio 3:
N: 2400
n: 200
k: 12 (2400/12)
Start p: 10
10, 10 + 12, 24, …, 10 + (200 – 1) * 12 => 10, 22, 24, …, 2398

Come usare excel?
=SEQUENZA (ampiezza campionaria; numero colonne; p; k)

Quindi con questi dati …
=SEQUENZA (200;1;10;12)

Il numero delle colonne è sempre 1, perché vogliamo una sola colonna di valori.



Il campionamento sistematico è problematico quando:
- **Lista non casuale o ordinata con criterio particolare**: se le persone nell’elenco sono messe secondo un criterio particolare (ad esempio: tutti gli studenti lavoratori sono vicini, prima tutti i maschi e poi tutte le femmine, prima chi ha voti alti e poi chi ha voti bassi), allora prendendo “uno ogni k” rischi di prendere sempre lo stesso tipo di persona.
- **Presenza di periodicità nella lista**: immagina che in una lista ogni 10° nome sia uno studente lavoratore oppure ogni 5° persona appartenga allo stesso reparto. Se il tuo passo di campionamento è proprio 10 (o 5), finirai per selezionare sempre quella stessa categoria.In poche parole, il campionamento sistematico funziona bene solo se la lista “mescolata e casuale”. Se invece la lista ha un ordine nascosto o una struttura ripetitiva, il metodo può scegliere sempre lo stesso tipo di individuo, e quindi il campione non sarà rappresentativo.

23/02

3. Campionamento stratificato: È la versione più organizzata e complessa del casuale semplice.
- Creazione degli strati: Il ricercatore divide arbitrariamente la popolazione in gruppi omogenei chiamati strati, basandosi su variabili come genere ed età (es. femmine 18-20 anni, maschi 18-20 anni, ecc.).
- L'estrazione: Da ogni singolo strato si estraggono le unità applicando il metodo casuale semplice.
- La proporzionalità: Per garantire la rappresentatività, il campione deve riflettere il "peso" reale che ogni strato ha nella popolazione. Se uno strato ha 900 persone e un altro ne ha 10, non si può estrarre un numero fisso (es. 5 unità per entrambi), altrimenti lo strato grande risulterebbe sottorappresentato. È necessario fare una proporzione basata sull'ampiezza totale del campione.

Quando scegliere un metodo o l'altro?
- Il Campionamento Casuale Semplice è più indicato quando l'obiettivo è l'esplorazione di un fenomeno nuovo.
- Il Campionamento Stratificato è preferibile quando l'obiettivo è la descrizione accurata di una popolazione di cui si conoscono già le caratteristiche principali.

4. **Campionamento a grappolo:** si suddivide la popolazione in gruppi eterogenei all'interno e si seleziona casualmente un grappolo. Nel campionamento a grappolo, la popolazione non viene selezionata direttamente individuo per individuo, ma viene prima suddivisa in gruppi chiamati *grappoli* o cluster.

   Questi grappoli devono essere eterogenei al loro interno, cioè contenere persone diverse tra loro (non tutte uguali).

   Successivamente si estraggono casualmente uno o più grappoli e si intervistano:
   - O tutte le persone che fanno parte dei grappoli selezionati
   - Oppure (in alcune varianti) solo una parte di esse

   Questo metodo è particolarmente utile quando la popolazione è molto ampia e distribuita sul territorio, perché permette di ridurre spostamenti, tempi e costi.

   Perché il quartiere conta? (Esempio sulla città): Immaginiamo di voler misurare il reddito familiare in una città. Supponiamo che:
   - La differenza tra quartieri sia 18 milioni
   - La differenza tra famiglie nello stesso quartiere sia 42 milioni

   La differenza totale sarà: 18 + 42 = 60 milioni

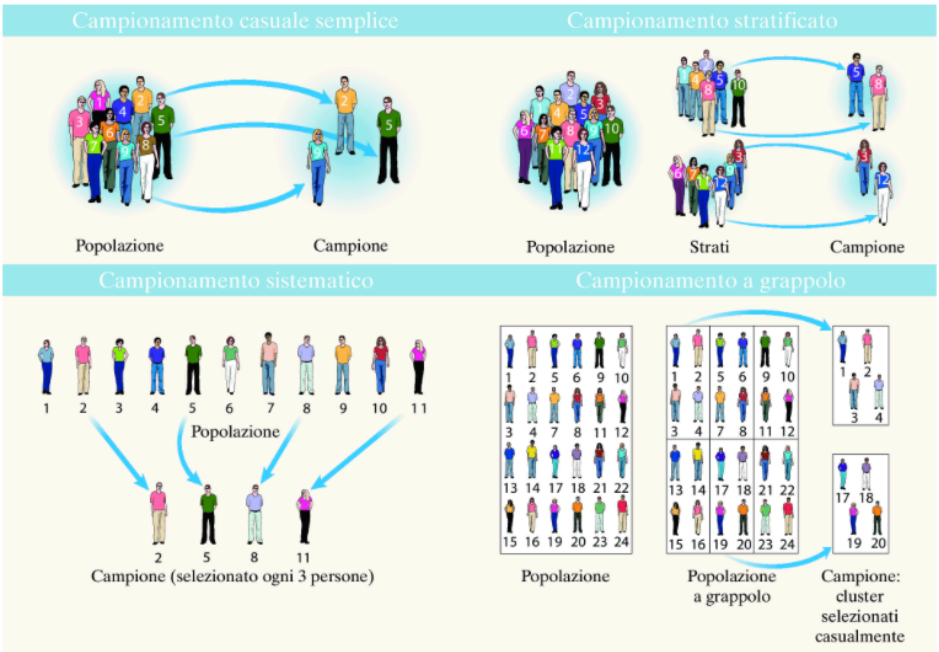
   Ora calcoliamo le percentuali:
   - 30% della differenza dipende dal quartiere (18/60 × 100)
   - 70% dipende dalle differenze tra famiglie nello stesso quartiere

   Questo significa che il quartiere influenza comunque il reddito, anche se la maggior parte della variabilità è interna.

   Esempio concreto:
   - Quartiere A: redditi tra 25.000 e 45.000 €
   - Quartiere B: redditi tra 60.000 e 80.000 €

   Dentro ogni quartiere c'è molta differenza (20.000 € di ampiezza), ma chi vive nel quartiere A è comunque mediamente più povero di chi vive nel quartiere B.

   Quindi il grappolo (quartiere) ha un peso. Nel campionamento a cluster si usa il concetto di **Design Effect**. Il Design Effect misura la perdita di efficienza rispetto al campionamento casuale semplice.



CAMPIONAMENTO CASUALE SEMPLICE	CAMPIONAMNETO A CLUSTER
Scegli 100 città diverse	Scegli 5 città
Intervisti 1 famiglia per città	Intervisti 20 famiglie in ciascuna
Devi viaggiare ovunque	Pochi spostamenti
Tempo e costi enormi	Costi molto più bassi + Risparmi logistici enormi

Perché c'è questa perdita? Perché le persone dentro lo stesso cluster non sono indipendenti: tendono ad assomigliarsi.
Se molte unità sono simili tra loro:
- Forniscono meno informazione nuova,
- Il campione è meno preciso rispetto a un campione casuale semplice della stessa dimensione

Il Design Effect dipende da:
- Quanto sono grandi i cluster
- Quanto sono simili le persone all'interno del cluster

Più sono simili => maggiore perdita di efficienza.
Quindi è utile usare il campionamento a cluster quando:
- La popolazione è geograficamente dispersa,
- I costi di rilevazione sono elevati,
- Non esiste una lista completa di tutte le unità,
- Si fanno indagini su larga scala,
- Serve semplificare la logistica.

Il cluster riduce i costi, ma perde un po' di precisione statistica.

Campionamento a due stadi
Il campionamento a due stadi è una variante del campionamento a grappolo.
La differenza fondamentale è questa:
- Nel cluster puro si selezionano i gruppi e si osservano tutte le unità al loro interno.
- Nel due stadi si selezionano prima i gruppi, e poi si estraggono solo alcune unità all'interno dei gruppi, con un'ulteriore estrazione casuale.

Quindi:
- Cluster = una fase
- Due stadi = due fasi

Come funziona?
1° stadio: si selezionano casualmente alcuni grappoli
2° stadio: dentro ciascun grappolo si selezionano casualmente alcune unità
Si usa quando costruire una lista completa di tutti gli individui è troppo difficile o troppo lungo.

Esempio semplice:
- Seleziono 5 quartieri
- In ogni quartiere seleziono 20 famiglie
- Non osservo tutte le famiglie del quartiere

Riduce i costi rispetto al cluster puro.

Esempio campionamento a due stadi:
Vuoi sapere quanto spendono in media le famiglie italiane per la spesa alimentare.
Non puoi intervistare tutte le famiglie d'Italia, quindi fai così:
1. **Primo stadio**: scegli casualmente 5 Comuni italiani (i comuni sono i tuoi "grappoli").
2. **Secondo stadio**: dentro ciascun Comune selezioni 20 famiglie a caso. Non tutte le famiglie, solo alcune.

Poi intervisti quelle 20 famiglie per ogni Comune.

Si chiama a due stadi perché ci sono due estrazioni consecutive:
- Prima scegli i gruppi (Comuni),
- Poi scegli le unità dentro i gruppi (famiglie).

Differenza con il cluster puro:
- Cluster puro = scelgo 5 Comuni e intervisto tutte le famiglie.
- Due stadi = scelgo 5 Comuni e dentro ognuno intervisto solo alcune famiglie.

Differenze tra metodi: quando confrontiamo i diversi metodi di campionamento, la differenza principale riguarda il livello di precisione statistica e la fattibilità pratica.

Il **campionamento casuale semplice** è quello che garantisce la massima precisione. In questo metodo ogni unità della popolazione ha la stessa probabilità di essere selezionata e le osservazioni sono indipendenti tra loro. Proprio per questo motivo fornisce stime molto affidabili e con errore statistico più basso. Tuttavia, è spesso difficile da applicare nelle indagini molto grandi, perché richiede una lista completa di tutte le unità e può comportare costi elevati.	Il **campionamento a due stadi** rappresenta una soluzione intermedia. La precisione è inferiore rispetto al casuale semplice, ma superiore rispetto al cluster puro. Questo accade perché, pur selezionando prima dei gruppi (ad esempio Comuni), al secondo stadio si estraggono casualmente le unità all'interno dei gruppi, recuperando parte della precisione persa. È quindi un buon compromesso tra rigore statistico e praticità operativa.	Il **campionamento a cluster puro**, invece, ha la minore precisione. Questo perché si selezionano interi gruppi e si osservano tutte le unità al loro interno, che spesso sono simili tra loro. La somiglianza interna riduce l'informazione effettiva raccolta e aumenta l'errore campionario. Tuttavia, nonostante la minore precisione, i metodi a cluster (puro o a due stadi) sono spesso più fattibili nella pratica. Permettono di ridurre costi, tempi e problemi logistici, soprattutto nelle indagini territoriali o nazionali.

In sintesi:
- Il casuale semplice è il più preciso ma il meno pratico.
- Il cluster è il meno preciso ma spesso il più economico e gestibile.
- Il due stadi è un compromesso tra precisione e fattibilità.

Conclusione:
Il campionamento probabilistico è il metodo più rigoroso dal punto di vista scientifico, perché elimina la soggettività e si basa su regole matematiche.

Proprio grazie all'**equiprobabilità** e all'**equirappresentatività**, permette di fare inferenze affidabili e di estendere i risultati dal campione all'intera popolazione con un margine di errore controllato.

24/02

Esercitazione 1: CAMPIONAMENTO
Campione casuale semplice
La selezione del campione casuale semplice può essere fatta secondo due modalità distinte:
1. **Con reinserimento**: l'unità selezionata viene reinserita nella popolazione e può essere estratta di nuovo
2. **Senza reinserimento**: l'unità selezionata viene rimossa dalla popolazione e non può più essere scelta —> quasi sempre

Nel campionamento casuale semplice, dalla popolazione estraiamo un campione.
Da una popolazione di N=10 unità, quanti campioni casuali di numerosità n=4 si possono estrarre?

N!/n!(N-n)! = 10!/4!6! = 5040/24 = 210
210 è il numero di campioni casuali, di numerosità 4, che si possono estrarre da una popolazione di 10 unità.

CAMPIONE STRATIFICATO
Un campione stratificato è ottenuto separando la popolazione in gruppi non sovrapposti chiamati strati.
Da ogni strato vengono poi estratte le unità con un campionamento casuale semplice.
Le unità all'interno di ogni strato dovrebbero essere omogenee rispetto a certe caratteristiche.

N=324
Due variabili: genere e tipo di sport
N=20

Come si trasforma il campionamento su Excel?
1. **Campionamento casuale semplice**: può essere fatto in 2 modalità distinte
   - Con reinserimento —> l'unità selezionata viene reinserita e può essere estratta di nuovo
   - Senza reinserimento —> l'unità selezionata viene rimossa dalla popolazione e non può più essere scelta (quasi sempre)

   Se la popolazione N = 10, quanti campioni casuali di numerosità n = 4 si possono estrarre?
   Si utilizza il coefficiente binomiale

$$\frac{N/n}{N!\,/\,n!\,(N-n)!} \qquad \frac{10!}{4!\,6!} \qquad \frac{10 * 9 * 8 * 7 * 6 * 5 * 4 * 3 * 2 * 1}{4 * 3 * 2 * 1 * 6 * 5 * 4 * 3 * 2 * 1} \qquad \frac{5040}{24} = 210$$

   210 COMBINAZIONI POSSIBILI

   ! = fattoriale = moltiplicazione del numero per tutti i numeri interi precedenti

   Come funziona su Excel?
   = COMBINAZIONE (10;4)
   = FATTORIALE (10) / FATTORIALE (4)* FATTORIALE (6)

   Per creare un frame è necessario disporre di una lista di campionamento
   Per estrarre un numero —> =CASUALE.TRA (minore;maggiore)

   Oppure —> formule, inserisci formule, cerca
   Per ripetere la funzione si trascina semplicemente il mouse verso il basso replicando la funzione

   Da un campione di 15 unità trovo la media generale della popolazione
   Errori di campionamento (es. numerosità troppo bassa rispetto all'ampiezza della popolazione, oppure errori dei non rispondenti, errori dei rispondenti)

2. **Campionamento stratificato**
   Separando la popolazione in gruppi chiamati strati, da ogni strato vengono estratte unità con il campionamento casuale semplice.
   Le unità all'interno di ogni strato dovrebbe essere omogeneo a certe caratteristiche

   In questo caso viene effettuata l'estrazione partecipanti alle olimpiadi divisi per genere e sport (2 strati: maschio e femmina, individuale e gruppo) per determinare l'età media del campione di 20 su 324.

		Team or Individual	
		Individual	Team
Gender	Female	132	34
	Male	109	49

Bisogna determinare il peso (la quota) per ogni strato:
Per calcolare il peso del primo strato faccio 132:324 = ….

STEP 1°: calcolo peso per ogni strato	
Peso 1° Strato	0,407407
Peso 2° Strato	0,104938
Peso 3° Strato	0,33642
Peso 4° Strato	0,151235

STEP 2°: determino unità da estrarre	
Unità 1° Strato	8,148148
Unità 2° Strato	2,098765
Unità 3° Strato	6,728395
Unità 4° Strato	3,024691

Oppure faccio = e clicco i due numeri di cui voglio calcolare il rapporto (per non scrivere i numeri)
Dopo aver calcolato il peso, ho la percentuale di ciascuno strato che deve essere calibrato sul campione che voglio ottenere
Quindi se il primo strato pesa 40%, devo moltiplicare 40% per 20 (numerosità del campione n)

Una volta ottenute le unità devo approssimare e assicurarmi che la somma delle unità sia 20

In questo caso 8 + 2 + 7 + 3 = 20

3. **Campionamento sistematico**
   Si ottiene un campione selezionando dalla popolazione un'unità ogni k (passo di campionamento) La prima unità va selezionata tramite estrazione da 1 a k e sarà p (start o punto di partenza)

   - Calcola N/n e arrotonda al numero più vicino
   - Questo valore è k = 25
   - Seleziono a caso uno start estratto con =CASUALE.TRA (1;25)

   Il campionamento a grappolo suddivide la popolazione N in gruppi eterogenei (grappoli), ciascuno dei quali rappresenta le caratteristiche della popolazione su scala ridotta.un campione a grappolo è ottenuto selezionando tutti gli individui all'interno del grappolo scelto in maniera casuale.
   I grappoli devono contenere unità eterogenee tra loro.

   Il campionamento a due stadi è una variante del campionamento a grappoli che si differenzia da questo poiché sostituisce la fase finale di rilevazione totale delle unità all'interno dei grappoli con un'estrazione casuale.
   Per cui, selezionerò al primo stadio un certo numero di grappoli con estrazione casuale e al secondo selezionerò n unità campionarie sempre attraverso un'estrazione casuale senza ripetizione.

2/03

VARIABILI, MODALITÀ, SCALE E FREQUENZE (pt. 1)
Ogni variabile ha dei modi di presentarsi che si chiamano **modalità**.

Auditel: per calcolare lo share di programmi come Sanremo, si parte dalla selezione di 16.000 famiglie tramite campionamento casuale, utilizzando liste elettorali o l'anagrafe comunale. In questo contesto, ogni aspetto misurato (sesso, età, titolo di studio) rappresenta una variabile, che in un foglio Excel corrisponde a una colonna.

Le Modalità e le Variabili Qualitative e Quantitative
Ogni variabile si manifesta attraverso diverse **modalità** (i modi di presentarsi del fenomeno). Ad esempio, la variabile sesso ha modalità come "maschio", "femmina" o "altro".
I fenomeni sono classificabili in due grandi famiglie,
1. **Variabili Qualitative:** descrivono qualità e si dividono in base a due scale di misura:
   - Scala Nominale: le modalità sono semplicemente un elenco e tra loro esiste solo un criterio di differenza o uguaglianza (es. colore dei capelli, abbonamenti TV o professioni).
   - Scala Ordinale: le modalità seguono una gerarchia o un ordine crescente/decrescente (es. titolo di studio, gradi militari, ruoli aziendali o livelli di soddisfazione del cliente).
2. **Variabili quantitative** esprimono quantità numeriche e si dividono in:
   - Discrete: le modalità sono numeri interi e finiti ovvero "senza virgole" (es. numero di figli, numero di macchine).
   - Continue: le modalità sono numeri con la virgola che esprimono variazioni infinitesimali (es. altezza, peso).

Esistono due tipi di scale di misura:

- Scala a Intervalli: lo zero non è assoluto ma convenzionale o arbitrario. La rilevazione non inizia dallo zero condiviso (es. la temperatura in Celsius che inizia da 35° sul termometro, le date storiche che hanno l'anno zero come convenzione, o le taglie dei vestiti).
- Scala di Rapporto: esiste uno zero assoluto e condiviso che indica l'assenza del fenomeno (es. il reddito).

Variabili qualitative		**Variabili quantitative**	
SCALA NOMINALE	SCALA ORDINALE	A INTERVALLI	DI RAPPORTO
Uguaglianza e differenza: le modalità sono solo un elenco non numerato	Ordine e gerarchia: esiste una gradualità o una relazione di "maggiore/minore"	Zero convenzionale: lo zero non indica assenza del fenomeno ma è un punto arbitrario	Zero assoluto: lo zero indica l'effettiva assenza della quantità misurata
Genere (M, F, A), Colore dei capelli o degli occhi, Professione, Abbonamenti Pay TV	Titolo di studio, Gradi militari, Classe socio-economica, Livelli di soddisfazione (Customer Satisfaction)	Temperatura Celsius (inizia da 35° sul termometro), Date storiche (anno zero convenzionale), Taglie dei vestiti	Reddito, Peso, Altezza, Distanza

Concetti Cardine: Variabili (X) e Modalità (x)

- **Variabile** (X): Si indica con la lettera maiuscola e rappresenta il carattere oggetto di studio (es. Genere, Età, Titolo di studio). In un foglio Excel, ogni variabile occupa una colonna.
- **Modalità** (x): Si indica con la lettera minuscola e rappresenta i modi specifici in cui la variabile si presenta (es. "Maschio", "Femmina").
- **Pedice** (i): Per generalizzare, usiamo il pedice i (la "i-esima" modalità).

Concetto di frequenza assoluta
La frequenza (ni) è il conteggio di quante volte una modalità si verifica nel collettivo:

- **N** indica la frequenza totale
- $\mathbf{n_i}$ indica la frequenza per ogni modalità x

GENERE	FREQUENZE (n_i)
FEMMINILE (x_1)	125 (n_1)
MASCHILE (x_2)	25 (n_2)
ALTRO (x_3)	10 (n_3)
TOTALE	160 (N)

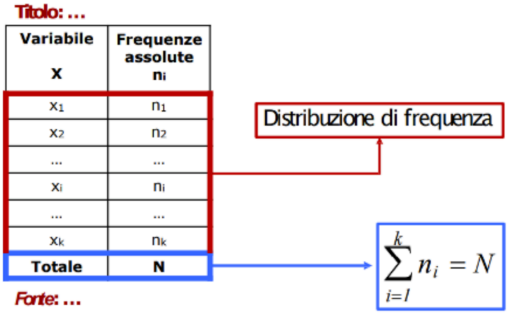
La Distribuzione di Frequenza
L'insieme di modalità e frequenze si chiama Distribuzione di Frequenza. È la prima forma di sintesi dei dati.

- Distribuzione Unitaria: Se ogni unità statistica ha una modalità diversa (tutte le frequenze sono pari a 1).
- Variabili in Classi: Quando i dati sono troppi (es. pesi continui), li raggruppiamo in intervalli (es. 45-50, 50-55).
- Il Dilemma del 50: Per convenzione, l'estremo inferiore è escluso. Nella classe 45-50, il valore 45 non c'è (si parte da 45,00...1), mentre il 50 è incluso. Questo serve a dare continuità: un valore come 50,99 deve "cadere" necessariamente in una classe senza dubbi o buchi.

Peso	**Frequenze**
45 - 50 (x_1 – x_2)	100 n_1
50 - 55 (x_2 – x_3)	25 n_2
55- 60 (x_3 - x_4)	35 n_3
Totale	160 N

Estremo inferiore classi escluso

Formalizzazione Matematica: La Sommatoria (∑)
Per non scrivere addizioni infinite utilizziamo la sommatoria

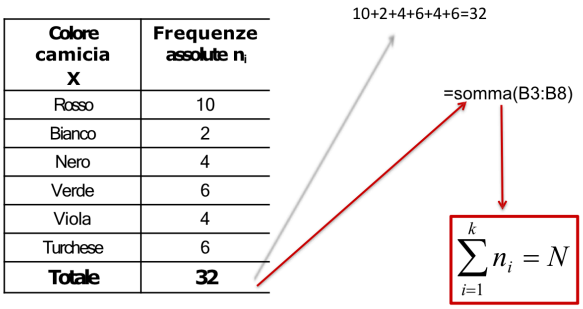


Si legge: "Sommatoria per i che varia da 1 a k di n con i". È lo strumento che permette di generalizzare le formule per migliaia di dati.

Sommatoria in Excel

1. Selezionare i dati (es. la colonna eta con i valori 20,21,22..
2. Vai su inserisci > tabella pivot
3. Trascina la variabile in righe
4. Trascina la stessa variabile in valori e imposta conteggio (per ottenere frequenze ni)

Esempio "Colore camicia" (Slide 16): Per sommare le frequenze (Rosso: 10, Bianco: 2, Nero: 4, Verde: 6, Viola: 4, Turchese: 6), in Excel useremo il comando: =SOMMA(B3:B8) → Risultato: 32.



3/03

VARIABILI, MODALITÀ, SCALE E FREQUENZE (pt. 2)

Variabile X	Frequenze assolute n_i
x_1	n_1
x_2	n_2
x_3	n_3
x_4	n_4
x_5	n_5
x_6	n_6
Totale	**N**

La **distribuzione di frequenze** è uno strumento che ci permette di organizzare e riassumere i dati in modo chiaro. Quando raccogliamo delle informazioni su un gruppo di persone (*per esempio studenti*), stiamo osservando una **variabile**, che si indica con la lettera maiuscola X. Le **modalità** con cui questa variabile si presenta si indicano con x_1, x_2, x_3… (lettera minuscola con un numero accanto).

Accanto a ogni modalità inseriamo la **frequenza assoluta**, che si indica con n_1, n_2, n_3…: essa rappresenta il numero di unità statistiche (*per esempio studenti*) che possiedono quella determinata caratteristica. Per calcolarla, semplicemente contiamo quante persone rientrano in ciascuna categoria.

La somma di tutte le frequenze assolute dà la **numerosità totale**, indicata con la lettera N maiuscola. L'ultima riga della tabella deve sempre riportare questo totale, perché rappresenta la somma di tutte le osservazioni.

È importante distinguere alcuni casi particolari:
Se tutte le frequenze assolute sono uguali a 1, ci troviamo di fronte a una **DISTRIBUZIONE UNITARIA:** significa che ogni modalità compare una sola volta. In questo caso, la numerosità totale N coincide semplicemente con il numero delle modalità osservate.

Diverso è il caso in cui tutte le frequenze siano uguali tra loro ma diverse da 1 (per esempio tutte pari a 10): qui non abbiamo una distribuzione unitaria, ma una **DISTRIBUZIONE UNIFORME** (o costante). Basta anche una sola frequenza diversa ≠ da 1 per non poter più parlare di distribuzione unitaria.

Variabile X	Frequenze assolute n_i
x_1	n_1
x_2	n_2
x_3	n_3
x_4	n_4
x_5	n_5
x_6	n_6
Totale	N

Colore Capelli	Frequenze assolute n_i
Biondo Cenere	1
Biondo caldo	1
Biondo scuro	1
Castano	1
Mogano	1
Nero	1
Totale	6

Distribuzione Unitaria

Colore Capelli	Frequenze assolute n_i
Biondo Cenere	10
Biondo caldo	15
Biondo scuro	10
Castano	25
Mogano	5
Nero	20
Totale	85

Oggi, per costruire queste tabelle, non è necessario contare tutto a mano. Possiamo utilizzare strumenti come Excel. Se abbiamo un elenco di dati (ad esempio una colonna con le età), possiamo creare una Tabella Pivot: selezioniamo i dati, andiamo su "Inserisci" e scegliamo "**Tabella Pivot**", poi trasciniamo nelle righe la variabile che vogliamo conteggiare. Excel calcolerà automaticamente le frequenze.
Attenzione: se abbiamo una colonna con numeri che rappresentano un codice identificativo (ID), quella non è una variabile numerica in senso statistico, ma una variabile nominale, perché i numeri servono solo come etichette e non hanno significato quantitativo.

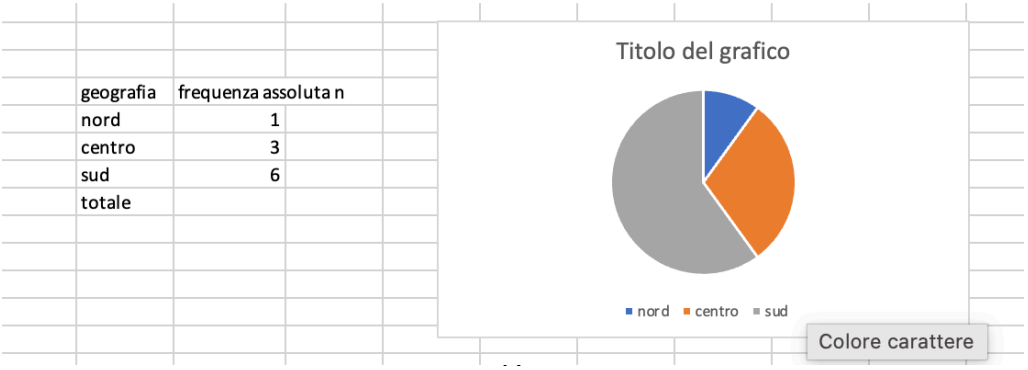
Tipi di distribuzione
Con una variabile nominale possiamo costruire tre tipi di distribuzione.
La prima è la **distribuzione di frequenze assolute**, di cui abbiamo già parlato.

La seconda è la **distribuzione di frequenze relative.** In questo caso vogliamo capire che PESO ha ciascuna modalità rispetto al totale. Si calcola facendo il rapporto tra la frequenza assoluta di una modalità e il totale N. La frequenza relativa si indica con f_i (lettera f minuscola con indice). Essa ci dice la quota o proporzione che quella categoria rappresenta sul totale.

$$f_i = \frac{n_i}{N}$$

La terza è la **distribuzione di frequenze percentuali.** Si ottiene moltiplicando la frequenza relativa per 100. In questo modo esprimiamo il peso di ciascuna modalità in percentuale.

Scala nominale (elenco) e scala ordinale (gerarchicamente)
Le variabili nominali sono misurate su **scala nominale**. Si tratta di caratteristiche che NON possono essere ordinate né misurate quantitativamente. Per esempio: il nome di una persona, il colore degli occhi, la nazionalità. Le modalità sono espresse con parole e non esiste un ordine gerarchico naturale tra di esse.



Esistono però variabili che, pur essendo espresse con categorie, possono essere ordinate. In questo caso parliamo di **scala ordinale**. Un esempio è il livello di istruzione: elementare, media, superiore, laurea. Sono categorie, ma hanno un ordine logico e gerarchico.

Anche per le variabili ordinali possiamo costruire la distribuzione di frequenze assolute, relative e percentuali. In più, poiché esiste un ordine, possiamo calcolare le **frequenze cumulate** (indicate con N_i).

Le frequenze cumulate si ottengono sommando progressivamente le frequenze assolute. La prima frequenza cumulata coincide con la prima frequenza assoluta. La seconda è data dalla somma della prima e della seconda frequenza assoluta, la terza è la somma delle prime tre, e così via. L'ultima frequenza cumulata **deve sempre coincidere con il totale N**. Se non coincide, significa che c'è un errore nei calcoli.

Livello di Istruzione	**Frequenze assolute n_i**
Elementare	3
Media	4
Lic. Superiore	3
Totale	**10**

Livello di Istruzione	**Frequenze relative f_i**
Elementare	0.3
Media	0.4
Lic. Superiore	0.3
Totale	**1**

Livello di Istruzione	**Frequenze percentuali p_i**
Elementare	30%
Media	40%
Lic. Superiore	30%
Totale	**100%**

Livello di Istruzione	**Frequenze cumulate N_i**
Elementare	3
Media	7
Lic. Superiore	10
Totale	**10**

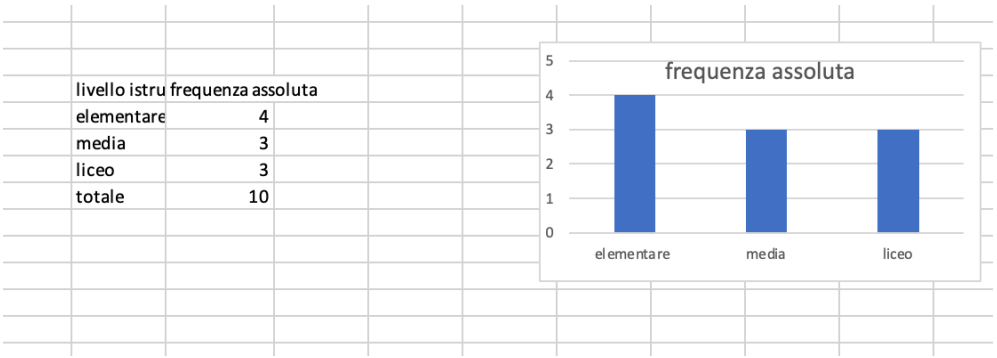
Si legge così:
- 7 persone hanno studiato almeno elementari e medie
- 10 persone hanno studiato almeno elementari, medie e liceo (in questo caso i valori sono 3 quindi coincide con il totale)

Le frequenze cumulate non si possono calcolare per le variabili nominali, perché manca un ordine gerarchico tra le modalità.

Rappresentazioni grafiche

Dopo aver costruito le tabelle, possiamo rappresentare i dati con un grafico.

- Per le variabili nominali, l'obiettivo è mostrare il PESO di ciascuna categoria sul totale. Il grafico più adatto è il **grafico a torta**, perché evidenzia visivamente le proporzioni. In Excel, si selezionano i dati (senza includere il totale), si va su "Inserisci" e si sceglie il grafico a torta, oppure si utilizza la voce "Grafici consigliati" e poi "Tutti i grafici".
- Per le variabili ordinali si utilizza invece il **diagramma a barre**. Anche in questo caso si selezionano i dati (escludendo il totale), si va su "Inserisci", poi "Grafici consigliati" e infine si sceglie il tipo di grafico desiderato.



È importante non confondere il diagramma a barre con l'istogramma: nell'istogramma le colonne sono attaccate, mentre nel diagramma a barre sono separate. Nel caso delle variabili ordinali (espresse in categorie), le barre devono rimanere separate.

DISTRIBUZIONE DI FREQUENZA

Le tabelle caratterizzate da modalità non aggregate sono:
- Distribuzione di frequenze

Le tabelle caratterizzate da modalità aggregate o congiunte sono:
- Distribuzioni doppie
- Distribuzioni di frequenze in classi

Se analizzo un fenomeno seguendo 2 classificazioni diverse (es. imprenditrici adulte - giovani e ripartizioni nord - ovest, nord- est) posso costruire una tabella univoca oppure dividere per età e per ripartizioni

La tipologia grafica più adatta è il diagramma a barre, ma per le tabelle a doppia entrata si utilizza la distribuzione percentuale.
Nell'asse orizzontale ci sono le territorialità e i colori differenziano il secondo fenomeno.
I colori rendono evidente la distribuzione in percentuale già a primo impatto.

La distribuzione di frequenza in classi avviene solo con dati quantitativi, raggruppando le singole modalità in classi —> intervalli numerici
Es. 0-10 anni
10-20 anni

La suddivisione in classi è opportuna quando:
- Variabili quantitative discrete (solo numeri interi)= se hanno molte modalità
- Variabili quantitative continue (numeri comprese le virgole) = sempre

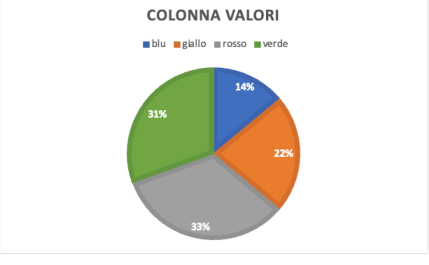
Le classi sono caratterizzate da:
- Limite inferiore (il valore più piccolo contenuto nella classe) sx
- Limite superiore (il valore più elevato contenuto nella classe) dx
- Ogni classe è un intervallo numerico che presenta un'ampiezza = differenza tra limite inferiore di una classe e il limite inferiore della classe successiva

Copia immagine LIMITI
Per rappresentare gli intervalli:
L'istogramma crea rettangoli attaccati tutti uguali tra loro, ciò che cambia è l'altezza dei rettangoli che scandisce la frequenza assoluta di ciascun intervallo

1. Quando le classi hanno la stessa modalità sarà uniforme (rettangoli di altezza uguale e quindi uguale frequenza)
2. Campandulare o simmetrica significa che gli estremi minori e superiori sono inferiori rispetto al centro
3. Quando i rettangoli più alti sono concentrati a sx dell'istogramma (valori più piccoli a sx) si chiama asimmetria positiva o a destra
4. Quando i rettangoli più alti sono concentrati a dx ossia la distribuzione sarà nei valori più alti si chiamerà asimmetria negativa o a sinistra

Etichette di riga	Conteggio di id
blu	5
giallo	8
rosso	12
verde	11
Totale complessivo	**36**

Colonna valori	Colonna valori
blu	5
giallo	8
rosso	12
verde	11
Totale	36



09/03

ISTAT: indicatori e indici e statistica per l'analisi delle dinamiche dell'universo femminile

La statistica è la scienza dello stato, ed è la scienza che dovrebbe regolare le nostre vite.

Democrazia e statistica:
La statistica è uno strumento fondamentale di democrazia;
La fondazione del **1926:** istituto centrale di statistica (ISTAT);
Dal **1946:** suffragio universale delle donne
ISTAT: nel 2026 festeggia 20 anni, è nato prima della democrazia

"Conoscere per deliberare" (Luigi Einaudi)

Non si può deliberare in un contesto che non si conosce. Dobbiamo conoscere per decidere.

"Il sistema statistico nazionale svolge un ruolo essenziale nel determinare il progresso economico e sociale del Paese".
(Sergio Mattarella, 2024)

I dati offerti sono strumenti per rendere più efficaci le politiche di intervento in settori che presentano squilibri o che necessitano di specifiche e dettagliate misure correttive al fine di incrementare il benessere collettivo a vantaggio delle generazioni presenti e future. Come faccio a fare un intervento se conosco la politica di intervento?

Potere VS Statistica:
"Paradosso degli aerei che non tornavano" (Spiegato durante la prima lezione)
La statistica cosa rappresenta? È la **rappresentatività** di tutto il collettivo di interesse, dobbiamo vedere i dati non di una porzione.
- Il rapporto tra potere e dati: la statistica permette al cittadino di basarsi su una verità oggettiva, rendendo il dibattito paritario.
- La statistica permette a una democrazia di decidere dove investire le risorse in modo efficiente, evitando sprechi basati su impressioni errate. La statistica è il sistema immunitario della democrazia.

La statistica funge da sistema immunitario di una democrazia:
1) **Responsabilità**: permette ai cittadini di misurare le promesse della politica si sono trasformate in risultati.
2) **Contrasto alle fake news:** in un'era di disinformazione, il dato ufficiale prodotto da istituti indipendenti è l'unico ancoraggio alla realtà.
3) **Inclusione**: una buona statistica rende visibili le minorazione o le fasce di popolazione vulnerabile, che altrimenti, rimarrebbero ai margini delle decisioni politiche
4) **Scelte consapevoli**: un cittadino informato su dati economici, sociali e ambientali può votare e partecipare basandosi su prove scientifiche

Diritto alla conoscenza: dobbiamo conoscere il contesto in cui ci troviamo
Se la conoscenza è un diritto allora l'accesso all'informazione è il prerequisito per tutti gli altri diritti.
- Se non conosco l'andamento del mercato del lavoro non posso orientare il mio futuro.
- Se non conosco le statistiche di genere non posso intervenire per contrastare le disuguaglianze.
- Se non conosco i miei diritti sanitari non posso curarmi.

"La statistica è la prima difesa democratica contro la tirannia dei pareri non fondati"

Un statistico del '900 dice che i problemi sociali sono accessibili e pensabili a chi conosce pochi ed elevati concetti, la soluzione a questi problemi è per pochi, così non abbiamo la capacità di conoscere o decidere. La statistica ha lo stesso peso di saper leggere e scrivere.
<u>Come sta la statistica ufficiale?</u>

Le aziende che ci offrono servizio sanno tutti di noi, soprattutto cosa siamo tentati a consumare.
- Io che faccio statistica ufficiale, come faccio a contrastare queste potenze?

***DOMANDA D'ESAME**

Variabile - Indicatore - Indice:

- **Variabile**: è una grandezza che varia, misura le variazione del spazio o del tempo. Può essere il numero di assunte in Lombardia.
- **Indicatore**: colui o ciò che indica; è un **rapporto** tra due quantità. Per esempio il tasso di occupazione femminile. Man mano che allarghiamo la nostra analisi alla complessità, aggreghiamo gli indicatori per arrivare agli indici.
- **Indice**: è qualcosa che è utile per indicare qualcos'altro ; ci sono più componenti che possono interagire tra di loro, per esempio l'indice di massa corporea (relazione tra peso e altezza di una persona), perché due variabili interagiscono tra di loro. Composizione di vari indici che fanno a forme un indicatore tra 0 e 100.

Indicatori sociali: negli anni 60 nascono questi ultimi; traggono origine dall'organismo sociale. misure statistiche quantitative, spesso organizzate nel progetto BES (Benessere Equo e Sostenibile), atte a descrivere aspetti non economici del benessere della popolazione, come salute, istruzione, relazioni sociali, sicurezza e ambiente, superando la sola visione basata sul PIL.

Granularità e tempestività:

ISTAT da un pò di tempo produce indici e indicatori fondamentali che cercano di arrivare al maggior dettaglio territoriale, sono indicatori granulari:

- Indice di fragilità comunale (IFC): fragilità del contesto comunale, quindi sociale o economica
- Indice di disagio socio economico a livello sub comunale
- Indice di povertà educativa
- Amisuradicomune: sistema di indicatori di genere in ogni regione; (indicatore di genere: rapporto tra due variabili, e il risultato ci darà la possibilità di leggere un pezzo in più dell'analisi)
- Previsione demografiche
- Indicatore del FRAME
- Indicatore territoriali per le politiche di sviluppo
- BEST

—>Indicatore di genere: confronto tra due variabili diverse

- Il livello di disagio: più il numero è alto, più disagio c'è; esempio fatto in classe in riferimento alle zone di Milano.

La statistica moderna dimostra che che non si tratta di una semplice coincidenza, ma di una correlazione strutturale. La fragilità La fragilità sociale e il disagio non colpiscono in modo "neutro": il genere funge spesso da moltiplicatore di vulnerabilità. Esiste un fenomeno che i sociologi chiamano "femminilizzazione della povertà" quale spiega come le donne siano statisticamente più esposte al rischio di esclusione sociale a causa di squilibri preesistenti.

Fragilità. Disagio, divari di genere:

1) **Divario economico:** (Gender Pay & Pension Gap). Le donne tendono ad avere carriere più frammentate (maternità, cura dei familiari) e sono sovrarappresentate in settori meno retribuiti. Risultato: Minore capacità di risparmio e, nel lungo periodo, pensioni più basse. Questo crea una "fragilità differita" che esplode in età anziana, dove le donne corrono un rischio di indigenza molto più alto degli uomini.
2) **Il carico di cura:** (Unpaid Work); La fragilità sociale è spesso legata alla mancanza di tempo. In Italia, la maggior parte del lavoro domestico e di assistenza (bambini e anziani) ricade sulle donne. Risultato: questo doppio lavoro riduce la possibilità di formazione, di carriera e di partecipazione alla vita pubblica, portando a un isolamento che è la base del disagio psichico e sociale.
3) **La violenza e dipendenza economica:** La disuguaglianza di genere si manifesta anche come controllo economico. Molte situazioni di disagio, profondo e marginalità derivano dall'impossibilità di lasciare contesti familiari violenti perché manca l'autonomia finanziaria.

La statica è una lente di ingrandimento: senza una statica di genere, queste dinamiche resterebbero invisibili. Se guardassimo solo il reddito medio pro capite di una nazione, non vedremmo che dietro quel numero si nasconde una distribuzione profondamente ingiusta che rende una metà della popolazione più fragile dell'altra.

"Le donne hanno sempre dovuto lottare doppiamente, hanno sempre dovuto portare due pesi, privato e sociale. Le donne sono la colonna vertebrale della società".

10/03

Concetti base della distribuzione di frequenza:

- **Frequenza Assoluta:** Conteggio del numero di osservazioni per ogni modalità.
- **Frequenza Relativa e Percentuale:** La relativa misura il peso di ogni modalità (numero tra 0 e 1); la percentuale (relativa x 100) è considerata la più comprensibile.

Le Frequenze Cumulate: Le **frequenze cumulate** rappresentano un modo per sintetizzare i dati indicando quante unità statistiche raggiungono "al massimo" un certo livello di una variabile.

La frequenza cumulata di una specifica modalità si ottiene sommando la sua frequenza assoluta a quelle di tutte le modalità che la precedono nell'elenco ordinato.

- **Esempio pratico:** Se abbiamo 3 persone con licenza elementare e 4 con licenza media, la frequenza cumulata della licenza media sarà **7** (ovvero 3 + 4).

Per quali variabili si possono usare?
Non si possono usare per tutto. La distribuzione delle frequenze cumulate ha senso solo per:

- **Variabili qualitative ordinali:** quelle che hanno un ordine logico, come il titolo di studio.
- **Tutte le variabili quantitative:** sia quelle **discrete** (numeri interi) che quelle **continue** (che richiedono uno strumento di misura).

Nota importante: Non possono essere fatte per le variabili **nominali** (come i colori: bianco, rosso, verde), perché non esiste un ordine che permetta di sommarle progressivamente.

Come interpretarle (Il concetto di "Al massimo")
Il valore della frequenza cumulata ci dice quante persone possiedono una caratteristica pari o inferiore a quella considerata.

- **Esempio del titolo di studio:** Dire che la frequenza cumulata per "licenza media" è 7 significa che **7 persone hanno al massimo la licenza media** (quindi includiamo sia chi ha la media, sia chi ha solo l'elementare).
- Allo stesso modo, se la frequenza cumulata per il "diploma superiore" è 10 (che corrisponde al totale del campione), significa che 10 persone hanno **al massimo** il diploma (comprendendo elementari, medie e superiori).

A cosa servono nella pratica?
Anche se ora sembrano solo un esercizio di calcolo, le frequenze cumulate sono indispensabili per calcolare in futuro indici statistici più avanzati, come:

- La **mediana** (il valore centrale della distribuzione).
- Il **primo e il terzo quartile** (che dividono la popolazione in quattro parti uguali).

In sintesi, le frequenze cumulate servono a "leggere" la popolazione in modo progressivo, aiutandoci a capire come si distribuiscono i soggetti man mano che la modalità della variabile cresce.

Possiamo rappresentarle con il diagramma o con le torte.

Quanti tipi di tabelle possiamo avere?

1. **Modalità non aggregate (viste fino ad ora)**
   - Distribuzioni di frequenze

2. **Modalità aggregate e modalità congiunte:**
   - Modalità Aggregate: Quando si raggruppano le modalità in classi o gruppi.
   - Modalità Congiunte (Distribuzione Doppia): Si osservano due caratteristiche dello stesso fenomeno contemporaneamente (tabella a doppia entrata).

- Esempio delle imprenditrici: Si analizza il fenomeno per età (giovane/adulta) e area geografica.
- Distribuzioni Marginali: Si ottengono sommando le righe (es. totale giovani/adulte) o le colonne (es. totale per area geografica) della tabella doppia.
- Prova del nove: La somma dei totali di riga deve sempre coincidere con la somma dei totali di colonna e con il totale generale (85 nell'esempio)
-

Distribuzione doppie & tabella a doppia entrata

Imprenditrici	Ripartizione	
	Nord **Ovest**	Nord Est
Giovani	15	20
Adulte	10	40
Totale	25	60

Esempio:
15 = ci sono 15 imprenditrici giovani che sono del Nord Ovest
Le distribuzioni congiunte si chiamano cosi perché possiamo costruire la distribuzione delle imprenditrici diverse per età o per ripartizione geografica .

Tabella età: 35 = Nord Ovest + Nord Est = 15 + 20 = 35
Tabella ripartizioni: 25 = 15 + 10 (colonna nord ovest)
60 = 20 + 40 (colonna nord est)

Prima distribuzione di frequenze

Imprenditrici	n_i
Giovani	35
Adulte	50
Totale	85

Seconda distribuzione di frequenze

Ripartizioni	n_i
Nord Ovest	25
Nord Est	60
Totale	85

Non si utilizza questa tipologia di diagramma a barre

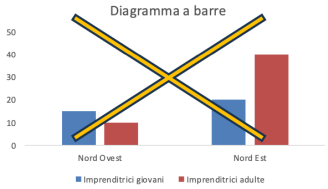


Diagramma a barre percentuali: Per le tabelle a doppia entrata preferibile usare barre percentuali invece di colonnine affiancate (foto sopra), perché mostrano immediatamente l'impatto e la differenza tra gruppi (es. Russia vs Ucraina nel 2022).

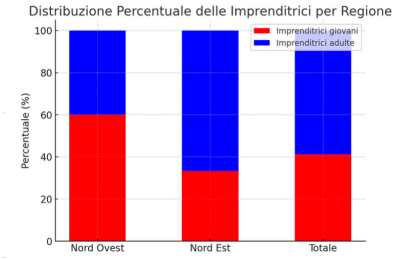
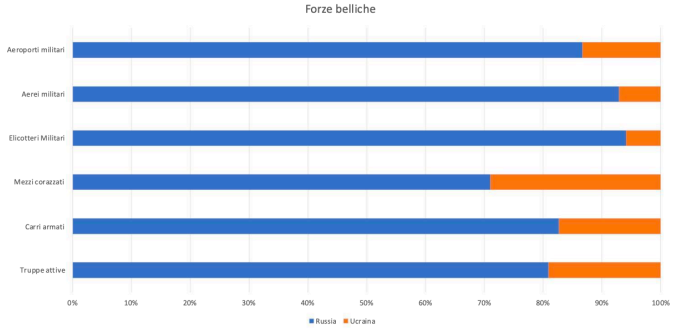


Diagramma a barre delle frequenze percentuali
Asse x orizzontale: caratteristica territoriale
Asse y verticale: percentuale dell'età delle imprenditrici

Esempio su dati reali: Distribuzione delle forze belliche (Guerra Russia-Ucraina)

Forze bellicihe	Russia	Ucraina
Truppe attive	850	200
Carri armati	12420	2596
Mezzi corazzati	30122	12303
Elicotteri Militari	544	34
Aerei militari	4173	318
Aeroporti militari	1218	187

Fonte: Infodata-Sole24ore, 2022



Per rendere intuitiva e veloce la lettura dei dati => diagramma a barre
Asse x: percentuale delle forze belliche
Asse y: caratteristica delle forze belliche

È immediato come la percentuale delle forze belliche russe sia nettamente superiore di quella ucraina

Dati reali del 2022 (tratti da Il Sole 24 Ore) relativi all'inizio del conflitto. La tabella originale elenca lo sbilanciamento delle forze belliche (truppe attive, carri armati, aerei, elicotteri) tra i due paesi. Ad esempio, la Russia disponeva di 850.000 truppe contro le 200.000 dell'Ucraina, e di oltre 12.000 carri armati contro circa 2.600.

- Il limite della tabella: Leggere un lungo elenco di numeri è faticoso e poco immediato per il pubblico. I numeri assoluti sono difficili da confrontare rapidamente per capire la reale entità del divario.
- La soluzione grafica: Un diagramma a barre percentuali (in questo caso ribaltato orizzontalmente) dove ogni barra rappresenta il 100% di una specifica forza bellica (es. il totale degli aerei militari disponibili tra i due paesi)

Esercitazione 2_Frequenze & Grafici
Tabella pivot: passaggi per la costruzione della tabella

1. Evidenzio tutta la tabella di cui voglio calcolare la tabella Pivot
2. Inserisci - tabella pivot
3. Pop up: dire che la tabella pivot deve essere nel foglio di lavoro esistente -> clicchi casella a caso dove vuoi posizionarla -> cliccare invio
4. Campi tabella pivot
   - Trascinare "colore" nello spazio “righe"
   - Trascino id nella colonna valori → schiacciare tasto "info" che appare di fianco a "id" → mettere "conteggio"|
5. Inserisci Inserisci sequenza Connessioni
6. Copiare tabella -> incollare speciale sotto, incollando solo valori
7. Abbiamo ottenuto una tabella di frequenza
8. Visto che il colore è una variabile qualitativa nominale → faccio grafico a torta
9. Seleziono tutti i dati della seconda tabella che ho creato→ NON SELEZIONO IL TOTALE
10. Inserisci → grafici consigliati → **torta**

Creazione della Tabella Pivot (Il "Conteggio")
La Tabella Pivot è descritta come la prima operazione fondamentale per contare le unità e creare tabelle di frequenza.

- Selezione dei dati: Si può selezionare la tabella con il mouse oppure cliccando su una cella qualsiasi e premendo contemporaneamente Ctrl + A (o Cmd + A su Mac) per evidenziare tutto automaticamente.
- Inserimento: Bisogna andare nel menù in alto su "Inserisci" e cliccare su "Tabella Pivot".
- Posizionamento: selezionare "Foglio di lavoro esistente" e cliccare su una cella vuota a fianco della tabella originale, per avere i risultati subito visibili senza cambiare foglio.
- Configurazione dei Campi:
  - Righe: Trascinare la variabile d'interesse (es. "Colore") nello spazio "Righe". Questo elencherà le modalità (blu, giallo, rosso, verde).
  - Valori: Trascinare la variabile "ID" nello spazio "Valori".
  - Importante: Excel di default potrebbe fare la "Somma" degli ID. Bisogna cliccare su "Somma di ID", andare su "Impostazioni campo valore" e selezionare "Conteggio" per ottenere la frequenza assoluta.

"Pulizia" dei Dati: Incolla Speciale
Una volta ottenuta la tabella pivot, non si deve lavorare direttamente su di essa perché è uno strumento che può "spostarsi".

- Procedura: Selezionare la tabella pivot, copiarla e incollarla poco più sotto usando "Incolla speciale > Valori" (l'icona con i numeri 1 2 3).
- Rinominare: Una volta incollati i valori "statici", è possibile cambiare manualmente le intestazioni, scrivendo ad esempio "Frequenza assoluta" al posto di "Conteggio di ID”.

Rappresentazione Grafica: La Torta
Per le variabili qualitative nominali (come il "Colore"), il grafico corretto da usare è la torta.

- Selezione: Selezionare i dati (modalità e frequenze), facendo attenzione a escludere il totale dalla selezione, altrimenti la torta verrebbe falsata.
- Inserimento: Andare su "Inserisci" > "Grafici consigliati" o "Tutti i grafici" e scegliere “Torta"
- Personalizzazione:
  - Percentuali: È meglio rappresentare le frequenze percentuali invece dei numeri assoluti, perché rendono più chiara la proporzione del fenomeno. Si può fare tramite il menù "Struttura grafico" o cliccando sul "+" del grafico (Elementi grafico) > "Etichette dati" > "Altre opzioni" e spuntando "Percentuale".

- Colori: Per coerenza visiva si consiglia di far coincidere il colore della fetta con il nome della modalità. Per farlo, bisogna fare doppio click sulla singola fetta, andare sull'icona del "Secchiello" (Riempimento) e cambiare il colore (es. blu per la modalità "blu").

Distribuzioni di frequenze in classi

Quando i dati sono **quantitativi**, la costruzione delle distribuzioni di frequenze si può fare raggruppando le singole modalità in **classi** => intervalli numerici

La suddivisione in classi è opportuna quando:

- **variabili quantitative discrete** => Ad esempio, se si studia il numero di magliette nell'armadio di un collezionista (che può averne 60) o l'età dell'intera popolazione italiana, non si può fare una tabella di centinaia di righe.
- **variabili quantitative continue** => Poiché richiedono uno strumento di misura (come bilancia o metro) e possono avere infinite virgole, le modalità singole sono quasi sempre diverse tra loro e vanno raggruppate

Le **classi** sono caratterizzate da:

- **Limite inferiore** => il più piccolo valore contenuto nella classe (sx)
- **Limite superiore** => il valore più elevato contenuto nella classe (dx)
- **Ampiezza** => differenza tra limite inferiore di una classe e il limite inferiore della classe successiva (limite inferiore - limite inferiore classe successiva)

Altezza (classi)
1,500 - 1,599
1,600 - 1,649
1,650 - 1,699

Le **classi** per variabili quantitative continue si possono rappresentare an che nel modo seguente:

Altezza (classi)
1,500 |- 1,600
1,600 |- 1,650
1,650 |- 1,700

Notazione	Estremo Inferiore	Estremo Superiore	Ampiezza Classe
$x_i - \| x_{i+1}$ $(x_i, x_{i+1}]$	*Escluso*	*Incluso*	$x_{i+1} - x_i$
$x_i \| - x_{i+1}$ $[x_i, x_{i+1})$	*Incluso*	*Escluso*	$x_{i+1} - x_i$

Esempio: Età

Età (classi)	n_i
25 - 34	12
35 - 44	21
45 - 54	7
55 - 64	6
65 - 74	9
75 e oltre	3
Totale	**58**

classe o tabella aperta

Esempio: Altezza

Altezza (classi)	n_i
1,500 - 1,599	3
1,600 - 1,649	14
1,650 - 1,699	18
1,700 - 1,749	11
1,750 - 1,799	7
1,800 - 1,899	4
1,900 e oltre	1
Totale	**58**

La convenzione di Inclusione ed Esclusione

Nelle tabelle capita spesso che il limite superiore di una classe coincida con quello inferiore della successiva (es. 1500-1600 e 1600-1650). Per capire dove inserire esattamente un valore (come il 1600), il corso segue questa convenzione:

- L'estremo inferiore (sinistra) è INCLUSO: Si indica con una stanghetta (|) o una parentesi quadra.
- L'estremo superiore (destra) è ESCLUSO: Si indica con un trattino (—) o una parentesi tonda.
- Esempio: Nella classe 1500-1600, il 1500 è contato, mentre il 1600 viene escluso e conteggiato nella classe successiva.

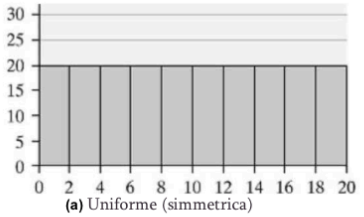
Classi Aperte

Esistono casi in cui l'ultima classe non ha un limite superiore definito per includere tutti i valori estremi, ad esempio "75 anni e più" o "1900 euro e oltre". In questo caso si parla di classe aperta o tabella aperta.

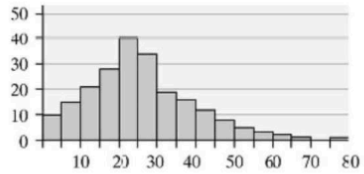
Rappresentazione Grafica: l'Istogramma
Per rappresentare le distribuzioni in classi si usa l'istogramma. A differenza del diagramma a barre, nell'istogramma i rettangoli sono attaccati tra loro (senza spazi) proprio per indicare la continuità degli intervalli numerici. L'altezza di ogni rettangolo rappresenta la frequenza di quella specifica classe

Per rappresentare gli intervalli:
L'istogramma crea rettangoli attaccati tutti uguali tra loro, ciò che cambia è l'altezza dei rettangoli che scandisce la frequenza assoluta di ciascun intervallo

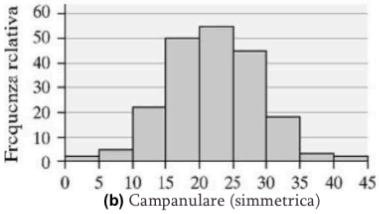
1. Quando le classi hanno la stessa modalità sarà uniforme (rettangoli di altezza uguale e quindi uguale frequenza)
2. Campandulare o simmetrica significa che gli estremi minori e superiori sono inferiori rispetto al centro
3. Quando i rettangoli più alti sono concentrati a sx dell'istogramma (valori più piccoli a sx) si chiama asimmetria positiva o a destra
4. Quando i rettangoli più alti sono concentrati a dx ossia la distribuzione sarà nei valori più alti si chiamerà asimmetria negativa o a sinistra



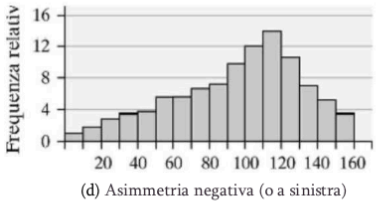
(a) Uniforme (simmetrica)



(c) Asimmetria positiva (o a destra)



(b) Campanulare (simmetrica)



(d) Asimmetria negativa (o a sinistra)

Non possiamo descrivere i dati qualitativi come uniformi, asimmetrici negativi o asimmetrici positivi.

1= importante riconoscere che i dati non sempre seguono un andamento perfettamente uguale a uno di quelli mostrati nella Figura 2.9. Per identificare la corretta forma della distribuzione é necessario essere un po' flessibili nell'analisi.

16/03

MEDIANA E QUARTILI - Misure di sintesi
Una misura di sintesi tradizionalmente utilizzata è la media, che però presenta un limite: non si traduce sempre in modo veritiero perché risente di boost, valori molto alti che modificano l'andamento regolare.
Quando si presentano modalità distanti e lontane dai valori costanti vengono dette anomale o outlier.
La mediana è definita media lasca e non risente dei valori anomali.

La famiglia delle misure di sintesi basate sulle posizioni è composta da:
- Primo quartile 25% Q1
- Secondo quartile o mediana (50%) Q2
- Terzo quartile 75% Q3

Se siamo interessati a descrivere il comportamento di parti del campione utilizziamo questi strumenti (es. quando lavoriamo in termini di percentuali; il primo 25% ha questo atteggiamento, la prima metà 50% ha questo atteggiamento, ecc.)

La mediana è una MISURA ROBUSTA perché NON è influenzata dai valori estremi mentre la media NON è robusta perché risente dei valori outlier

Quindi → Non si può sovrapporre il concetto di media e mediana perché riportano a misure differenti.

Un valore sintetico descrive parzialmente la variabile:
- **Media** aritmetica: variabili quantitative
- **Moda**: variabili quantitative e qualitative
- **Mediana** e **quartili**: variabili quantitative e ordinali

Il concetto di frequenza cumulata sottende che le frequenze coincidano con il concetto di posizione: le prime 100 tacche corrispondono ad un grado e corrispondono alle prime 100 posizioni.
→ La **frequenza** si può vedere come posizione.

La posizione si può esprimere in termini percentuali: alla posizione che si trova al 50% corrisponde un determinato valore - modalità (es. un valore per esprimere l'altezza).
La mediana divide in due parti uguali il mio collettivo e mi permette di far coincidere la posizione con la frequenza cumulata (che è serie ordinata progressiva).

La mediana di una variabile è il valore che occupa la posizione centrale e si può trovare sia nelle numerosità pari che dispari. Nelle numerosità pari per determinare la modalità che corrisponde alla posizione mediana avverrà il fenomeno di interpolazione.

I percentili dividono una distribuzione ordinata in 100 parti uguali. Si lavora sulle posizioni delle modalità che corrispondo alle frequenze.
Es. il 25% degli studenti ha avuto 30 all'esame di statistica

- 25esimo percentile corrisponde a 25/100 = 1/4
- 50esimo corrisponde a 50/100 = 1/2
- 75esimo corrisponde a 75/100 = 3/4

$i = 1 + (n-1) * p$ (percentile) — calcolo della posizione del campione
$i = 1 + (N-1) * p$ — calcolo della posizione della popolazione

N maiuscolo o minuscolo va a indicare il totale delle frequenze, perciò valgono entrambe le formule, è sufficiente prendere il totale (non cambia)

Voglio costruire la mediana che corrisponde al percentile 0,5

I quartili hanno dei percentili definiti:
- Q1 = 0,25
- Q2 = 0,5
- Q3 = 0,75

Faccio → i = 1 + (7-1) 0,5 = 4 → perciò prendo in considerazione la posizione 4 e osservo a quale variabile x1 corrisponde

In questo caso la mediana è 21 → 21/30 è il voto che crea due sottogruppi di uguale numerosità
Devo sempre creare la colonna delle frequenze cumulate

Se voglio calcolare Q1, faccio → i = 1 (7-1) 0,25 = 2,5
La posizione 2,5 ricade tra 2 e 3, ma seguendo il metodo dell'interpolazione:
Il 2,5 si compone di una parte intera 2 e una parte decimale 0,5 e si trova tra due modalità che sono 19 (x1) e 20 (x2)
Q1 = x1 + 0,5 (x2 - x1) → formula computazionale generalizzata iesima
Q1 = 19 + 0,5 (20-19) = 19 + 0,5 (1) = 19,5

Variabile X: voto Statistica

x_i	n_i	Ni
19	2	2
20	1	3
21	1	4
22	1	5
23	1	6
25	1	7
Totale	7	

i=1+(7-1)0,5=
i=1+(6*0,5)=
i=1+3= **4**

ESERCIZIO 1

xi (LIBRI LETTI)	ni (frequenze)	NI (freq. cumulate)
0	3	3
1	4	7
2	5	12
3	2	14
4	1	15
Totale	15	

La tabella riporta il numero dei libri letti in un mese da un gruppo di amici. Determina quanti libri ha letto il 50% delle persone, sapendo che tale valore corrisponde alla mediana della distribuzione.

Applichiamo la formula per il calcolo della mediana che individua la POSIZIONE della mediana.
Formula: $i = 1 + (15 - 1)\ 0{,}5 \rightarrow i = 1 + 7 = 8$
Dove si trova la mediana? Nella posizione 12 della tabella di frequenze cumulate.
Per cui la modalità che corrisponde alla posizione 12 nelle frequenze cumulate è 2 → almeno il 50% degli alunni hanno letto 2 libri

NB!
Non è necessario calcolare la posizione della mediana con il fenomeno dell'interpolazione perché la posizione superiore alla tabella è 7 che esclude 8. Quindi tutti i valori tra 8 e 12 compreso corrispondono alla posizione 12. L'interpolazione ricorre quando, nel calcolo dei quartili, risulta un numero con una parte decimale.

SU EXCEL:
Non si può usare la funzione mediana senza trascrivere tutti i valori in elenco perché devo trasformare la DISTRIBUZIONE DI FREQUENZE in DISTRIBUZIONE UNITARIA.
Quindi compilo l'elenco ripetendo la modalità tante volte quante si presenta (in base alla sua frequenza)
Es. ripeto la modalità zero 0 per 3 volte … ecc
Dopodiché calcolo la mediana

xi	ni	Ni		xi		
0	3	3		0		
1	4	7		0		2
2	5	12		0		
3	2	14		1		
4	1	15		1		
totale (n)	15			1		
				1		
i = 1+(n-1)*0,5				2		
i =1+(15-1)*0,5				2		
i = 1+ (14*0,5)				2		
i = 1 + 7 = 8				2		
				2		
				3		
				3		
				4		

ESERCIZIO 2

La tabella riporta il numero di assenza mensili registrate da 7 dipendenti di un'azienda. Determinare primo quartile, secondo e terzo sapendo che corrispondono al 25 percentile, 50 e 75 della distribuzione.

xi	ni	Ni (cumulate)
2	1	1
3	1	2
6	1	3
8	1	4
10	1	5
12	1	6
14	1	7
Totale	7	

Al contrario della precedente, questa tabella presenta una distribuzione unitaria perché tutte le modalità si presentano 1 sola volta.

Mediana (che si calcola per prima) → i = 1 (6*0,5) = 1+3 = 4 → posizione 4 della frequenza cumulata che corrisponde alla modalità 8 (almeno 50% hanno fatto 8 assenze)
La mediana è pari a 8

Per calcolare il primo quartile → Q_1: i = 1 + (6*0,25) = 2,5 quindi cerco la posizione del primo quartile tra 2 e 3 applicando l'INTERPOLAZIONE
Perché? Perché la posizione del quartile non è intera e non coincide esattamente con un'osservazione ma si ottiene interpolando, ossia determinando il valore compreso tra le due osservazioni consecutive.
Osservazioni x_2 e x_3
$X_2 = 3$
$X_3 = 6$
Parte decimale = 0,5

Formula → $Q_1 = 3 + 0,5$ (parte decimale di 2,5) * (x3 - x2)
$= 3 + 0,5 * (6-3) = 3 + (0,5 * 3) = 3 + 1,5 = 4,5$

Devo sommare a X_2 quello che gli manca ossia la parte decimale 0,5

Calcoliamo il terzo quartile → i = 1 + (6 * 0,75) = 5,5 corrisponde alla POSIZIONE del terzo quartile
Devo nuovamente ricorrere all'interpolazione perché il terzo quartile ha una parte decimale. Le osservazioni saranno 10 e 12.
$Q_3 = 10 + 0,5 (12-10) = 10 + (0,5*2) = 11$

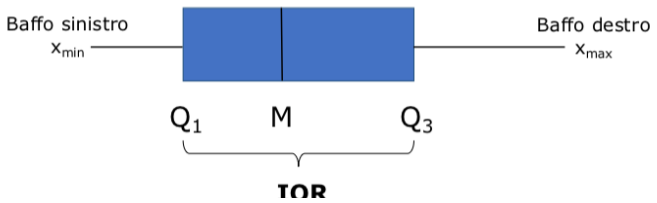
Come si calcola il quartile su EXCEL?
= QUARTILE (PRIMO VALORE NUMERICO - in questo caso 2 - : ULTIMO VALORE NUMERICO - 14) ; quartile → a seconda del quartile che voglio calcolare si apre la tendina con le diverse opzioni:

- 0 valore minimo
- 1 primo quartile
- 2 mediana
- 3 terzo quartile
- 4 valore massimo

IMPORTANTE!
- Tra il primo numero e l'ultimo ci sono i DUE PUNTI :
- Tra l'ultimo numero e il quartile c'è PUNTO E VIRGOLA ;

xi		
2		
3	Q1	4,5
6		
8	Q2	8
10		
12	Q3	11
14		

BOXPLOT
Il Boxplot o diagramma a scatola e baffi è un grafico che consente di rappresentare una distribuzione mediante 5 numeri → sintesi dei 5 numeri



IQR = differenza interquatile

Il BOXPLOT ci aiuta nella forma della distribuzione: quando è simmetrica, significa che osservazioni (modalità) sono bilanciate, ossia distribuite in modo regolare intorno al centro. In una distribuzione perfettamente simmetrica

Media = mediana = moda

La lunghezza dei baffi è rappresentata da:
- Baffo sinistro → si traccia una linea da Q_1 fino al limite inferiore LI → minimo
- Baffo destro → si tracia una linea da Q_2 fino al limite superiore LS → massimo

Il boxplot è costruito come una scatola che va dal primo quartile Q_1 al terzo quartile Q_3.
La linea interna rappresenta la mediana e la x all'interno della box indica la media (che può non coincidere con la mediana in caso di ASIMMETRIA.

Su EXCEL:
1. Si riporta l'elenco delle modalità xi
2. Si identificano i valori minimo (zero) e massimo (quattro)
3. Si calcolano i quartili (prima la mediana, poi Q_1, infine Q_3)
4. Si costruisce il grafico sotto dicitura "SCATOLA E BAFFI"

In base alla forma del boxplot si interpretano i dati e la simmetria/asimmetria.
Solitamente i dati non sono dispersi e vi è poca variabilità tra dati piccoli e grandi
Quando la x della media si sovrappone sulla linea della mediana significa che la distribuzione è simmetrica.

Simmetria → Q_2 è equidistante da Q_1 e Q_3 (verde)

Se ciò non accade, ci troviamo davanti ad una distribuzione asimmetrica, che può essere:
- A destra se Q_2 è più vicino a Q_1 rispetto che a Q_3 (blu)
- A sinistra se Q_2 è più vicino a Q_3 rispetto che a Q_1 (arancione)

Le asimmetrie servono ad indicare i dati anomali o outlier.

30/03

Individuare i Valori Anomali

Ci concentriamo su come "catturare" e visualizzare i **valori anomali** (outlier), ovvero quelle valori che si distanziano in modo "poderoso" dal cuore dei dati.

L'artificio dei limiti artificiali: Per capire se un valore è davvero anomalo, non ci si può affidare solo all'osservazione visiva. Bisogna creare due **limiti artificiali** (uno inferiore e uno superiore) che definiscono i confini della "normalità" della distribuzione.

Creare nuovi limiti artificiali significa svolgere un calcolo. Se sono interessato a scoprire se esistono valori anomali più piccoli o più grandi devo costruire un limite inferiore e uno superiore. Successivamente confronto il limite inferiore con x_{min}

Se $\mathbf{x_{min} < }$ **limite inferiore** = allora è anomalo

Esempio:

- Il limite inferiore (LI) risulta essere pari a 20
- Nella distribuzione dei dati (ad esempio i voti di un esame), è presente una modalità pari a 12
- Poiché il valore 12 è più piccolo del limite inferiore di 20, esso "supera" il confine della normalità verso il basso
- Il valore 12 viene classificato come valore anomalo

Nella *rappresentazione grafica*, questo 12 verrebbe indicato posizionando una stellina al di là della barra del baffo inferiore per segnalare visivamente l'anomalia. Questo metodo serve a non confondersi: anche se un valore è il minimo della distribuzione (x min), non è detto che sia un outlier finché non viene confrontato con il limite artificiale.

Le Formule per calcolare i limiti artificiali:

Limite Inferiore (LI): Q_1 - 1,5 x IQR.

- Si parte da Q_1 perché l'obiettivo è verificare se esistono modalità "molto più basse" rispetto al primo 25% della distribuzione. Sottraendo a Q1 una quota fissa di variabilità (1,5×IQR), stabiliamo un confine oltre il quale i valori sono considerati troppo distanti per essere normali

Limite Superiore (LS): Q_3 + 1,5 x IQR.

- Si parte da Q3 perché si vuole osservare se c'è qualcosa di "molto più grande" rispetto al 75%della popolazione. In questo caso, aggiungiamo la quota di variabilità a Q3 per definire il limite superiore

La costante 1,5:

Si tratta di una **costante empirica** proposta da Tukey (l'ideatore del boxplot). Attraverso diverse simulazioni, Tukey individuò che 1,5 è la misura che meglio riesce a intercettare i valori anomali senza scambiare per anomala una modalità che è solo vicina alla distribuzione.

La differenza interquartile IQR:

Si calcola *matematicamente* sottraendo il valore del primo quartile Q1 dal valore del terzo quartile Q3, seguendo la formula: IQR = Q_3 - Q_1

In termini *interpretativi*, l'IQR è una misura di dispersione che indica dove si concentra il 50% centrale dei dati, ovvero la parte della popolazione compresa tra il 25° e il 75° percentile. *Visivamente*, all'interno di un grafico boxplot (diagramma a scatola e baffi), l'IQR corrisponde esattamente all'altezza o alla lunghezza della scatola centrale.

Regola decisionale:

- Se il valore minimo (x_{min}) è **più piccolo** del LI, allora è valore anomalo
- Se il valore massimo (x_{max}) è **più grande** del LS, allora è valore anomalo

Rappresentazione grafica: I valori anomali vengono indicati sul grafico con una **stellina** o un punto al di là dei baffi.
ESEMPIO per mostrare un caso in cui non c'è un valore anomalo nonostante la presenza di un x_{min}:

L'esempio numerico dei limiti
Partendo da una distribuzione con i seguenti valori:

- Q_1 (Primo quartile): 19
- Q_3 (Terzo quartile): 36
- IQR (Differenza interquartile): 36 - 19 = 17
- Quota di variabilità: 1,5 x 17 = 25,5

Sulla base di questi dati, vengono costruiti i due limiti:

- Limite Inferiore (LI): 19 - 25,5 = -6,5
- Limite Superiore (LS): 36 + 25,5 = 61,5

Caso 1: Assenza di valore anomalo (x_{min})
Ipotizziamo di avere un valore minimo nella distribuzione pari a **4** quindi $x_{min} = 4$.

- **Confronto:** Si confronta l' x_{min} (4) con il limite inferiore (-6,5).
- **Risultato:** Poiché **4 è più grande di -6,5** (si trova cioè a destra del limite sulla retta dei numeri), il valore non supera il confine dell'anomalia.
- **Conclusione**: In questo caso non abbiamo valori anomali nella parte bassa della distribuzione, anche se 4 è il valore più piccolo registrato.

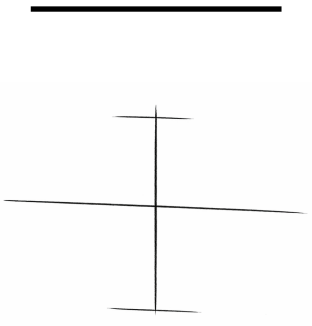
Caso 2: Presenza di valore anomalo (x_{max})
Nello stesso esempio, si ipotizza un valore massimo pari a **83** ($x_{max} = 83$).

- **Confronto:** Si confronta l'x_{max}(83) con il limite superiore (61,5).
- **Risultato:** Poiché **83 è superiore a 61,5**, esso "va nella fascia dei valori anomali".
- **Conclusione:** L'83, pur essendo l'ultima modalità della distribuzione, celava un'identità di **valore anomalo** e va segnalato con una **stellina** sul boxplot.

Questo dimostra che l' x_{min} e l'x_{max} possono essere "confondenti": solo il calcolo dei limiti artificiali permette di decifrare con precisione se siamo in presenza di un comportamento realmente anomalo o semplicemente di un valore basso/alto ma ancora nella norma.

Casi Particolari di Boxplot
Alcune varianti grafiche che possono capitare all'esame:

- **Boxplot come una linea:** Se in un grafico vedi solo una linea orizzontale (o verticale) senza la "scatola", significa che la distribuzione è **uniforme** (tutte le modalità sono pari allo stesso valore, es. studenti della stessa età). In questo caso, le modalità sono tutte uguali: Q_1, Q_2 e Q_3 coincidono nello stesso punto e la differenza interquartile (IQR) è pari a **zero. Non è necessario calcolare mediana e quartili.**
- **Baffi senza scatola:** Se i baffi esistono ma la scatola è una linea, significa che il valore minimo e massimo sono diversi, ma il 50% centrale dei dati (i quartili) coincide.



Introduzione alla Media Aritmetica
La Moda
La **moda** è definita come la modalità che presenta la **frequenza più elevata**.

- **Versatilità:** È l'unica misura di sintesi che può essere applicata a **tutti i tipi di variabili**, sia qualitative (anche nominali, come il colore dei capelli) che quantitative.
- **Significato:** Indica semplicemente ci che accade più spesso nel collettivo analizzato.

La Mediana
La **mediana (Q_2 o 50° percentile)** è il valore che occupa la **posizione centrale** in una distribuzione ordinata, dividendo il collettivo in due gruppi di uguale numerosità.

- **Robustezza:** È definita una **"misura robusta"** perché **non è influenzata dai valori anomali** (outlier). Ad esempio, se in un gruppo di persone con voti bassi si aggiunge un "30 e lode", la mediana non cambierà drasticamente, a differenza della media.
- **Applicabilità:** Si può calcolare solo per variabili **quantitative** e **qualitative ordinali**.
- **Calcolo:** Richiede necessariamente la creazione della colonna delle **frequenze cumulate (Ni)**, che permettono di far coincidere la frequenza con la posizione dei soggetti.

La Media Aritmetica
La media aritmetica è un valore di sintesi riferito a una variabile quantitativa e fornisce un valore di equilibrio all'interno della distribuzione di frequenze.
La "Cultura del Dato" e Trilussa: Esempio della famosa poesia del "Pollo" di Trilussa per spiegare il limite della media: se io mangio due polli e tu zero, in media ne abbiamo mangiato uno a testa, ma la realtà è ben diversa.
Il limite della media: La media **non è una misura robusta** perché è fortemente influenzata dai valori anomali. Da sola non è una buona sintesi: deve essere sempre accompagnata da una misura di **variabilità** (che verrà spiegata nella lezione successiva) per capire quanto i dati oscillino attorno ad essa.
Simbologia:

- **$\hat{x}$ (x soprassegnata)**: indica la media del **campione**
- **μ** (mu): indica la media della **popolazione**.

$$\bar{x} = \frac{\sum_i^k x_i n_i}{n} \qquad \mu = \frac{\sum_i^k x_i n_i}{N}$$

Calcolo della Media
Vengono illustrati i passaggi per calcolare la media in diverse situazioni:
1. Distribuzione Unitaria: Si sommano tutti i valori e si divide per il numero di unità (n).
2. Distribuzione di Frequenze Assolute: È l'approccio più comune.

- **Step 1:** Creare una colonna dove si moltiplica ogni modalità (x_i) per la sua frequenza (n_i).
- **Step 2:** Fare la **sommatoria** di questi prodotti $\Sigma\ x_i \times n_i$
- **Step 3:** Dividere il totale ottenuto per la numerosità totale (n).

Proprietà Fondamentali della Media
Tre concetti "da evidenziatore" per l'esame:

1. Internalità: Il valore medio deve essere sempre compreso tra il valore minimo e il valore massimo della distribuzione.
2. Mancanza di robustezza: La media risente degli outlier (a differenza della mediana).
3. Somma degli scarti: La somma degli scarti dalla media è sempre uguale a zero.

Poiché gli scarti si compensano, per misurare la distanza reale dei dati dalla media sarà necessario in futuro elevare questi scarti al quadrato.
Scarti = differenza tra modalità e media

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Uso di Excel
Per calcolare la media su Excel nel caso di un vettore di dati, si usa la funzione:

=MEDIA(intervallo_dati).

Se invece si ha una tabella di frequenza, bisogna procedere manualmente creando la colonna dei prodotti e dividendo poi per il totale, come spiegato nel calcolo teorico.

- Media: =MEDIA(intervallo) .
- Mediana: =MEDIANA(intervallo) (richiede di trasformare i dati in una distribuzione unitaria).
- Quartili: =QUARTILE(intervallo; quarto) (dove 1 è Q_1, 2 è la Mediana e 3 è Q_3).

Esercizio per calcolare media e per dimostrare che la differenza delle modalità e la media sono pari a zero.
Ragionare sulla formula* applicarla e verificarne la veridicità

Media aritmetica $\Sigma(x_i - media)\, n_i = 0$

xi	ni = frequenz	xi * ni	xi - media	(xi - media)*ni
18	10	180	-6,5319149	-65,319149
20	12	240	-4,5319149	-54,382979
25	50	1250	0,4680851	23,404255
28	12	336	3,4680851	41,617021
30	10	300	5,4680851	54,680851
totale	94	2306		0
media		24,531915		

Osservando la formula predispongo tot colonne in base ai calcoli che devo fare

1. Calcolo della media: $\mu=\frac{\sum_i^k x_i n_i}{N}$ = ((18x10)+(20x12)+(25x50)+(28x12)+(30x10)):94=24,53
   - Su excel: =MEDIA(intervallo_dati), non lo posso applicare alla colonna x_i perchè non tiene conto della frequenza
   - Vanno riportati tutti i decimali se no la media non risulta corretta

2. Verifico che la sommatoria dia come risultato 0

Il limite della Media: una misura incompleta
Sebbene fornisca un'idea del fenomeno, viene definita una **misura incompleta** perché non permette di comprendere il contesto di riferimento.

- **Esempio del pranzo:** Se cinque persone spendono in media 20,60€, si potrebbe pensare a un pranzo uniforme. In realtà, potrebbero esserci persone che hanno speso pochissimo e altre moltissimo: la sola media "nasconde" queste differenze.
- **La necessità della Variabilità:** Per descrivere davvero un fenomeno, bisogna sapere come i dati "oscillano" o variano rispetto al valore medio.

Il concetto di Variabilità e Strategia
Conoscere la variabilità ha risvolti pratici fondamentali:

- **Alta Variabilità:** Se l'intervallo di variabilità intorno alla media è molto ampio, significa che i comportamenti sono molto diversi; in questo caso, è necessario adottare **strategie differenti** per i diversi gruppi.
- **Bassa Variabilità:** Se i dati sono tutti vicini alla media (intervallo stretto), la popolazione è omogenea e una **singola strategia** può andare bene per tutti.

Oscillazione: I consumi e le scelte variano rispetto al valore medio di riferimento => se dovessi utilizzare soltanto il valore medio per descrivere un fenomeno perderei la caratteristica di variabilità.
Per poter arrivare ad un **intervallo di variabilità,** entro cui gli estremi si colloca il valore medio**.** Più l'intervallo è ampio più i comportamenti e le scelte sono variabili e bisogna adottare diverse strategie all'interno dell'intervallo.

Es. media dei voti in un gruppo universitario
N (num componenti gruppo)

12							
13				scarto	scarto ^2		
14	xi	ni	xi*ni	xi - media	xi - media ^2	xi - media ^2 * ni	
15	18	3	54	-5,8	33,64	100,92	
16	20	2	40	-3,8	14,44	28,88	
17	28	3	84	4,2	17,64	52,92	
18	30	2	60	6,2	38,44	76,88	
19	totale	10	238			259,6	
20	MEDIA		23,8				
21	DEVIANZA					259,6	
22	VARIANZA	devianza : 10				25,96	
23	SCARTO QUADRATICO MED					5,09509568	
24	I DUE ESTREMI		MEDIA + SCARTO		28,8950957		
25			MEDIA - SCARTO		18,7049043		

Passaggi matematici per calcolare la variabilità:
Step 1: Calcolo della Media
Si moltiplica ogni modalità per la sua frequenza, si somma il tutto e si divide per il totale delle unità.

- **Risultato:** Il voto medio ottenuto è **23,8**.

Step 2: Calcolo della Devianza
Per capire quanto i voti si allontanano dalla media, non si può usare la semplice somma degli scarti perché, per proprietà della media, la loro somma sarebbe sempre **zero**. Bisogna ricorrere al concetto geometrico di **distanza**, elevando gli scarti al quadrato.

1. Scarto al quadrato
2. **Ponderazione:** Si moltiplica ogni scarto al quadrato per la sua frequenza
3. **Sommatoria:** La somma di questi valori è la **Devianza** (numeratore della varianza), che nell'esercizio è pari a **259,6**.

Step 3: Varianza e Scarto Quadratico Medio

- **Varianza:** Si divide la devianza per la numerosità 259,6 / 10 = 25,96. Questo valore è però espresso in "trentesimi al quadrato", una scala difficile da interpretare.
- **Scarto Quadratico Medio :** Si estrae la radice quadrata della varianza per tornare alla scala originale.
- **Risultato** 5,095

Step 4: Definizione degli Estremi
L'intervallo si ottiene sottraendo e aggiungendo lo scarto quadratico medio alla media:

- **Limite inferiore:** 23,8 - 5,095 = 18,7.
- **Limite superiore:** 23,8 + 5,095 = 28,89.
- **Interpretazione:** Poiché la "forchetta" tra 18,7 e 28,9 è molto ampia, concludiamo che nel gruppo c'è **tanta variabilità** e i voti non sono affatto omogenei.

Popolazione vs Campione: La Correzione
Differenza fondamentale nella nomenclatura e nel calcolo a seconda del collettivo analizzato:

Caratteristica	**Popolazione**	**Campione**
Simbolo Varianza	(Sigma quadro)	S^2
Simbolo Deviazione	(Scarto Quadratico Medio)	(Deviazione Standard)
Denominatore	Si divide per **N**	Si divide per **n - 1**

La divisione per **n-1** nel campione è una **correzione** necessaria perché la media campionaria può oscillare rispetto a quella della popolazione; questa correzione rende lo stimatore della varianza "corretto".

Casi Particolari e Excel

- **Variabilità Zero:** Se σ^2 o S^2 sono pari a **0**, significa che tutte le unità hanno la stessa modalità. La distribuzione è **uniforme** (piatta) e il **boxplot** corrispondente appare come una semplice **linea**.
- **Excel:** La funzione rapida per la media è `=MEDIA(intervallo)`. Per la radice quadrata si può usare `=RADQ(valore)` oppure elevare il numero a `^0,5` (che equivale a 1/2).

	Varianza	Scarto quadratico medio (sqm)
Popolazione	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$	$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$
Campione	**Varianza campionaria** $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	**Deviazione standard** $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$

$$\sigma^2 = \frac{\sum_{i=1}^{k}(x_i - \mu)^2 n_i}{N}$$

POPOLAZIONE

$$s^2 = \frac{\sum_{i=1}^{k}(x_i - \bar{x})^2 n_i}{n-1}$$

CAMPIONE

In assenza di variabilità Varianza e varianza campionario valgono 0.

13/04

Progetto Mela: Metriche, Entertainment, Legale e AI (Lezione domanda Bonus)
L'evoluzione della musica: ci sono tante musicalità simili ma in realtà possono non essere state realizzate in modo naturale.

Il seminario, che vede la partecipazione dell'avvocato Renato Maraschi, si pone l'obiettivo di studiare il mercato della discografia e della musica attraverso l'applicazione di metriche statistiche per fare proiezioni e previsioni.

- **L'evoluzione del mercato:** Si analizza il cambiamento avvenuto in 14 anni, confrontando il 2012 (anno del successo di Adele con *Skyfall*) con il 2026.
- **Ibridazione:** Oggi assistiamo a un passaggio dalla pura creatività umana a una **generazione ibrida**, dove naturale e artificiale si fondono nella scrittura di testi, arrangiamenti e campionamenti.

L'Impatto Economico e il Ruolo delle Metriche
La statistica non è uno studio sterile, ma il motore che permette alle grandi **Major** (Warner, Universal, Sony) di decidere le proprie strategie.

- **Numeri del mercato:** Il settore musicale ha numeri imponenti: una valorizzazione prevista di **36,13 miliardi di dollari nel 2026**, con una crescita annua del **35%**.
- **Previsioni e Cluster:** Attraverso i modelli statistici (come i grafici a dispersione o *scatter plot* che verranno approfonditi), le etichette individuano i **cluster di ascoltatori**, prevedono il successo di un brano e decidono dove investire il budget di marketing.
- **La spinta asiatica:** Si prevede che l'area Asia-Pacifica trainerà il mercato fino al 2031, cambiando i gusti globali attraverso il consumo mobile.

Aspetti Legali e Diritto d'Autore nell'Era dell'IA
L'avvocato Maraschi sottolinea il divario tra la legge sul diritto d'autore (risalente al **1941**) e la realtà tecnologica attuale.

- Input vs Output:
  - **Input:** I dati e i brani inseriti nelle piattaforme per "istruire" l'IA. Spesso, sottoscrivendo i termini d'uso dei social, cediamo diritti senza accorgercene.
  - **Output:** Il risultato generato. Se l'IA crea interamente un'opera (es. "fammi una canzone triste"), sorgono problemi sulla titolarità dei diritti e sul blocco dei metadati per violazione.
- **Diritto Morale vs Sfruttamento Economico:** È fondamentale distinguere tra la paternità dell'opera (chi l'ha creata) e chi può guadagnarci.
- **Il Plagio:** Nel mondo dell'IA, il confine tra ispirazione e plagio (tradizionalmente fissato su 8 battute) diventa molto labile.

Il Caso degli Influencer e le Nuove Regole

Il seminario tocca anche il mondo dei **micro-influencer** e degli influencer (citando il caso Ferragni e Adriana Spink).

- **Sanzioni e Agicom:** È intervenuta l'Agicom per definire le categorie e le regole di trasparenza (advertising, hashtag).
- **Inquadramento contributivo:** Un tema caldo al Ministero riguarda il versamento dell'INPS (ex ENPALS) per gli influencer, essenziale per garantire la loro futura pensione.

Sintesi e "Domanda Bonus": Creatività e Criticità

La conclusione della lezione offre gli spunti chiave per la "cultura del dato":

- **Creatività come Valore:** Nonostante l'IA possa generare musica velocemente, il talento e la creatività umana rimangono l'unico bene capace di fare la differenza sul mercato a lungo termine.
- **Pensiero Critico:** Lo studente di statistica deve essere critico: non deve solo leggere i dati, ma chiedersi **come sono stati misurati** e se la sintesi proposta è corretta.
- **Blockchain e NFT:** Viene accennato il ruolo dei *Non-Fungible Tokens* per garantire l'unicità e l'autenticità di un'opera d'arte digitale in una catena sicura.

Punto chiave per l'esame: Ricorda che la media da sola non basta; serve sempre la variabilità, e nel mercato dell'entertainment, le metriche servono a trasformare la creatività in un modello di business solido e tutelato.

14/04

Introduzione alla correlazione

La statistica bivariata si riferisce all'insieme delle statistiche che studiano 2 diversi fenomeni/ caratteristiche osservate nella stessa collettività. Le due variabili si chiameranno x e y.
È interessante vedere se si sviluppano relazioni su due variabili e come si muovono.
Può esserci relazione tra variabili quantitative e qualitative. Per ognuna di queste relazioni esiste una misurazione specifica. Le scale di misura dei due fenomeni possono avere punteggi minimi e massimi diversi.

La relazione può prendere il nome di correlazione: non implica necessariamente una causa, ossia non significa che una variabile provoca l'altra, ma indica semplicemente che esiste un legame nel modo in cui variano insieme.

Se non si tiene conto della differenza delle scale di misura e della dispersione dei dati si avranno errori nel confronto delle variabili.

Per risolvere questo problema, la formula della correlazione trasforma le unità in z-score (punteggio standardizzato)
Uno **z-score** rappresenta la distanza di ciascun valore dalla media in unità di deviazioni standard.

Popolazione

$$z = \frac{x_i - \mu}{\sigma_i}$$

Campione

$$z = \frac{x_i - \bar{x}}{s_i}$$

- x → valore osservato
- M → media dei dati
- Sigma = deviazione standard per osservare quanto i dati sono dispersi

1. Il primo passaggio è togliere l'influenza della media (rumore) calcolando lo SCARTO:

   x - media

   Questo ci dice se un valore è sopra/sotto la media e di quanto, ma da solo non basta, perché non tiene conto della variabilità

2. Nel secondo passaggio si divide per la DEVIAZIONE STANDARD e questo serve a rendere i dati confrontabili tra loro, perché si elimina l'influenza della dispersione dei valori:

$$\frac{x - \text{media}}{\sigma}$$

In questo modo otteniamo un valore normalizzato, che non ci dice solo se siamo sotto o sopra la media, ma anche di quanto in termini relativi.

Lo z-score ha alcune proprietà fondamentali:
- Ha media 0
- Ha varianza 1
- Non ha unità di misura, è confrontabile tra variabili diverse

Interpetazione:
- Se z > 0 il valore è sopra la media
- Se z < 0 il valore è sotto la media
- Se z = 0 il valore coincide con la media

Più grande è il valore (in positivo o negativo), più è lontano dalla media

Esempio
Immaginiamo di avere 6 studenti che hanno fatto 2 test diversi di ingresso per l'università. Se guardiamo solo i punteggi grezzi, possiamo pensare che lo studente 5 sia migliore perché ha ottenuto il punteggio più alto. Tuttavia, questo può essere ingannevole, perché: non sappiamo quanto fosse difficile il test né quanto i punteggi siano distribuiti.
Calcolando gli z-score per ciascun test, trasformiamo i risultati in una scala comune. A questo punto confrontiamo le prestazioni in modo corretto.

$\mu_y = 164.17$ $\mu_x = 63.00$

$\sigma_y = 5.34$ $\sigma_y = 5.69$

id	y_i	x_i	$y_i - \mu_y$	$x_i - \mu_x$	z_y	z_x
1	160	55	-4,17	-8,00	-0,78	-1,41
2	155	62	-9,17	-1,00	-1,72	-0,18
3	165	65	0,83	2,00	**0,16**	**0,35**
4	170	68	5,83	5,00	1,09	0,88
5	170	71	5,83	8,00	1,09	1,41
6	165	57	0,83	-6,00	0,16	-1,06

$$z_3^y = \frac{165 - 164{,}17}{5{,}34}$$

$$z_3^x = \frac{65 - 63}{5{,}69}$$

Come leggere la formula sulla dx: z-score riferito alla terza unità della variabile y.

Se lo studente 5 ha z-score alto in entrambi i test, significa che è sopra la media in entrambi e lo è in misura significativa rispetto agli altri. → Quindi possiamo dire con sicurezza che è il migliore.

Una volta che i dati sono standardizzati (trasformati in z-score) possiamo studiare la correlazione tra variabili, rendendo tutte le osservazioni osservabili sulla stessa scala.
Per visualizzare meglio i risultati si ricorre ai grafici a barre affiancate, dove per ogni individuo si mostrano gli z-score dei due test. Si confortano così le prestazioni, per capire se coincidono con la media o la superano/non la raggiungono.

IN EXCEL:
- Per trovare il valore massimo degli z-score puoi usare =MAX (intervallo n.)
- Per trovare il valore minimo =MIN (intervallo n.)
- Per il costruire il grafico basta selezionare SOLO le colonne degli z-score e inserire grafico a barre
- Per calcolare il numeratore si deve sottrarre la variabile x/y alla media: la media va bloccata:

$MEDIA$ → esempio C10

Alternativa: scrivere il valore numerico corrispondente alla media

Altre funzionalità dello z-score

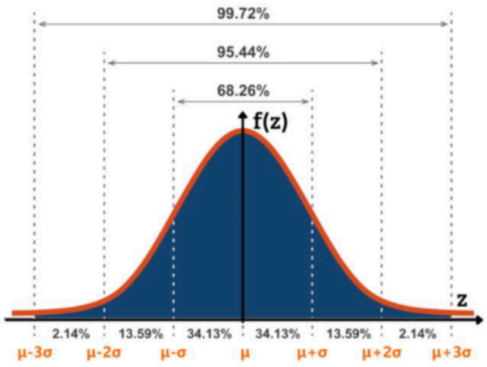
La trasformazione z-score che elimina effetto della media e della variabilità è molto importante in statistica. Quando si parla di media e moda, in realtà stiamo descrivendo un comportamento ordinario.

Immaginando la disposizione dio un mucchio di farina versato sul tavolo, potremmo dire che nella forma alta di un mucchio si concentra la normalità, nei lati la dispersione. Se dobbiamo costruire regole per generalizzare il comportamento sociale facciamo riferimento al centro. Se si dovesse disegnare questa forma, si otterrebbe l'andamento a campana → distribuzione principe-normale che generalizza il comportamento in statistica. La maggior parte dei dati si concentra vicino alla media, mentre pochi valori sono lontani.

La normalità viene modellata sulla base dei valori di media e scarto quadratico medio.
La curva normale standardizzata è definita da due parametri fondamentali:
- La media indica il centro della distribuzione
- La deviazione standard indica quanto i dati siano dispersi

Lo z-score trasforma qualsiasi distribuzione in una distribuzione standardizzata, facendo sì che la nuova media diventi pari a zero e la deviazione standard pari a 1.
Una volta standardizzata la distribuzione, possiamo intercettare delle affermazioni generali per generalizzare il comportamento:

Se il fenomeno segue la distribuzione standardizzata, si avranno i seguenti risultati chiamati INTERVALLI TIPICI che si calcolano a partire dalla media:
- tra μ ± 1σ → cade circa il 68% delle osservazioni della popolazione
- tra μ ± 2σ → cade circa il 95% delle osservazioni
- tra μ ± 3σ → cade circa il 99% delle osservazioni



• Se un valore è dentro ±1σ → è molto normale
• Se è dentro ±2σ → è ancora abbastanza normale
• Se è oltre ±3σ → è molto raro

20/04

Correlazione lineare, scatterplot e coefficiente di correlazione

1. Due variabili quantitative e correlazione

Quando abbiamo a che fare con due variabili quantitative siamo interessati a vedere come si muovono l'una nei confronti dell'altra, per capire se esiste una correlazione. Per esempio: *se in prossimità della prova costume diminuiscono le calorie assunte, c'è anche una diminuzione del peso?*

Sono tutte dinamiche che coinvolgono variabili quantitative misurate sullo stesso soggetto. Ci chiediamo quindi che tipo di interdipendenza ci sia tra le variabili: può esserci concordanza oppure discordanza. In questa parte lavoriamo solo in contesti di linearità: ci fermiamo quindi allo studio della correlazione lineare.

Distinzione importante!

- **Correlazione lineare**: relazione tra due variabili quantitative che può essere rappresentata e studiata tramite una tendenza lineare.

- **Correlazione spuria**: correlazione "non pura", cioè apparente, senza un collegamento razionale tra le variabili. Esempi: all'aumentare del numero di cicogne rosa aumenta anche il numero di neonati; le operazioni di Nicolas Cage sembrano collegate al numero di morti per annegamento in piscina.

Le correlazioni spurie coinvolgono due variabili che appartengono alla stessa unità di analisi, ma non hanno un legame realistico o razionale. Non sono quindi informazioni adatte a descrivere una vera relazione di interdipendenza.
Quando usiamo la statistica dobbiamo costruire una storia che abbia senso e che serva a dimostrare qualcosa. Devono esserci credibilità e realismo; quando questi mancano, si rischia di amplificare il mondo delle fake news.

2. Rappresentazione grafica: scatterplot

Anche nel caso di due variabili possiamo rappresentare i dati in maniera grafica. Si utilizza lo **scatterplot**, cioè il diagramma di dispersione.
Lo scatterplot si ottiene facendo riferimento alle informazioni di base del piano cartesiano: ogni punto è rappresentato da due coordinate, una sull'asse X e una sull'asse Y.

- La prima variabile caricata viene chiamata **X** oppure **xi**.
- La seconda variabile viene chiamata **Y** oppure **yi**.
- Nella rappresentazione grafica, la **X** rappresenta le ascisse, mentre la **Y** rappresenta le ordinate.

Le diverse modalità diventano quindi le nostre coordinate. Se ci fosse una terza variabile, potrebbe essere indicata come **Z**.
Esempio: se dobbiamo rappresentare Anna sul piano cartesiano, Anna avrà coordinate **900;50**.

Rappresentazione dimetrica

Questa rappresentazione grafica si chiama **dimetrica**, perché sulle ascisse lavoriamo con una scala di misura, per esempio le calorie, mentre sulle ordinate lavoriamo con un'altra scala di misura, per esempio il peso. In questo modo mettiamo congiuntamente scale di misura differenti.
Le modalità delle variabili possono quindi essere disegnate sullo scatterplot.
Dallo scatterplot, anche se non avessimo davanti le tabelle di calorie e peso, potremmo ricavare i valori rappresentati. In alcuni esercizi potrebbe capitare di dover completare la tabella dello scarto
standardizzato partendo proprio dallo scatterplot.

3. Configurazioni dello scatterplot

Lo scatterplot può assumere diverse configurazioni. Possiamo avere un addensamento:

- **Lineare crescente**: dal basso verso l'alto. Le due variabili tendono ad aumentare entrambe; siamo in un contesto di concordanza.
- **Lineare decrescente**: dall'alto verso il basso. Siamo in un contesto di discordanza: una variabile aumenta mentre l'altra diminuisce.
- **Non lineare**: esiste una forma di interdipendenza, ma non assume una forma di linearità. C'è una relazione, ma non la sappiamo quantificare con una retta.
- **Assenza di correlazione lineare**: i punti non mostrano una tendenza lineare riconoscibile.

Fare lo scatterplot è importante perché dalla rappresentazione grafica possiamo già formulare ipotesi sulle interdipendenze tra le variabili. Tuttavia dire soltanto che c'è concordanza, oppure che all'aumentare di una variabile aumenta anche l'altra, dà poca informazione: per questo è necessario quantificare la forza della relazione.

4. Coefficiente di correlazione

Per rispondere alla domanda "quanto è forte questa interdipendenza?" facciamo riferimento alla misura del **coefficiente di correlazione**, che aiuta a capire la forza e l'intensità della relazione tra due variabili.

Il coefficiente di correlazione nasce nei primi decenni del Novecento ed è noto come coefficiente di **Prive-Pearson.** Gli studiosi dell'epoca volevano quantificare queste misure e, in un famoso simposio, venne presentata una misura simile da studiosi che non si conoscevano ma arrivarono alla stessa conclusione. Correlazione e regressione, insieme a media e varianza,

sono il fulcro delle tipologie bivariate. La correlazione è la misura dell'intensità delle relazioni tra le interdipendenze, con X e Y sullo stesso piano.

Il coefficiente si indica con la lettera **r** minuscola.

$$\rho(X,Y) = \frac{\sum_{i=1}^{N}\left(\frac{x_i - \mu_x}{\sigma_x}\right)\left(\frac{y_i - \mu_y}{\sigma_y}\right)}{N} = \frac{\sum_{i=1}^{N} z_i^x z_i^Y}{N}$$

Quando r = +1, c'è una perfetta relazione lineare positiva.

Per interpretare al meglio una nuvola di punti bisogna fare riferimento all'immagine ipotetica di una retta. Più ci allontaniamo dal numero 1, cioè dalla bisettrice ideale, più i punti si spostano e diventano sparsi nel grafico. Man mano che ci avviciniamo verso lo zero, anche dalla parte negativa, il coefficiente diventa sempre più debole.

Esempio: calorie e peso
Nell'esempio riportato nelle foto, la variabile X corrisponde alle calorie e la variabile Y corrisponde al peso. Per costruire il grafico su Excel bisogna inserire prima X e poi Y.
Per controllare i calcoli su Excel si può usare la formula:

ID	X: calorie	Y: peso
1	929	66
2	1035	61
3	1366	64
4	1317	71
5	919	62

=CORRELAZIONE(intervallo X; intervallo Y)

Esempio mostrato nella foto: =CORRELAZIONE(C4:C8;D4:D8).
Attenzione: per fare correttamente la rappresentazione grafica su Excel bisogna mettere prima la colonna X e poi la colonna Y.

5. Scarto standardizzato
Per calcolare la correlazione si lavora sugli scarti standardizzati. Lo scarto standardizzato di X si ottiene sottraendo la media e dividendo per il sigma, cioè per lo scarto quadratico medio/ deviazione standard.

z = (valore - media) / sigma

Nel procedimento mostrato a lezione, prima va fatto il prodotto tra gli scarti standardizzati e poi la somma.

Esempio: durezza del terreno e tempo di perforazione
Ci chiediamo se la durezza del terreno è in correlazione con il tempo impiegato a bucare. I dati riportati nella slide si basano su uno studio geologico.
X rappresenta la profondità alla quale ha inizio la perforazione; Y rappresenta il tempo impiegato, in minuti. L'esercizio richiede di disegnare un diagramma di dispersione dei dati, cioè uno scatterplot.

Domande da porsi per verificare la presenza di correlazione
- Le due variabili aumentano entrambe?
- Le due variabili diminuiscono entrambe?
- Una variabile aumenta e l'altra diminuisce?

Con queste domande iniziamo a leggere tutti i dati e a capire la direzione della relazione.

Calcolo della correlazione con scarti standardizzati
Nella slide sul calcolo della correlazione con scarti standardizzati vengono riportati: X, Y, scarto

standardizzato di X, scarto standardizzato di Y e prodotto tra gli scarti standardizzati. Si sommano poi i prodotti ottenuti e si divide per il numero delle osservazioni.

Il valore finale indicato nella slide è circa **0,773**. Questo segnala una correlazione positiva abbastanza forte tra le due variabili dell'esempio.

Terreno	X	Y	Scarto stand. X	Scarto stand. Y	Prodotto
1	35	5,88	-1,825	-1,479	2,699
2	50	5,99	-1,525	-1,332	2,031
3	75	6,74	-1,025	-0,329	0,337
4	95	6,10	-0,625	-1,185	0,741
5	120	7,47	-0,125	0,648	-0,081
6	130	6,93	0,075	-0,075	-0,006
7	145	6,42	0,375	-0,757	-0,284
8	155	7,97	0,575	1,317	0,757
9	160	7,92	0,675	1,250	0,843
10	175	7,62	0,975	0,848	0,827
11	185	6,89	1,175	-0,128	-0,151
12	190	7,90	1,275	1,223	1,559
Somma	1515	83,83			9,274
Media	126,250	6,986			
s.q.m. (s)	50,005	0,748			

Riassunto operativo
- Individuare le due variabili quantitative X e Y.
- Costruire lo scatterplot, inserendo prima X e poi Y.
- Osservare la forma della nuvola di punti: crescente, decrescente, non lineare o assenza di correlazione.
- Calcolare gli scarti standardizzati sottraendo la media e dividendo per sigma.
- Moltiplicare gli scarti standardizzati di X e Y.
- Sommarli e ricavare il coefficiente di correlazione.
- Interpretare r: vicino a +1 relazione positiva forte; vicino a -1 relazione negativa forte; vicino a 0 relazione debole o assente.

21/04

Riepilogo Teorico: Il Coefficiente di Correlazione (r o ρ)
Coefficiente di correlazione lineare: è la misura dell'intensità e della direzione della relazione lineare tra due variabili quantitative.

Si indica con $\rho_{x,y}$ (o r) ed è calcolato come:

$$: r = \frac{\sum (z_x \cdot z_y)}{N}$$

dove si fa il prodotto tra gli z-score delle due variabili per ogni osservazione, si sommano questi prodotti e si divide per il numero totale delle osservazioni.

Quando si studia un fenomeno che coinvolge due variabili, queste possono:
- muoversi nella stessa direzione (si "attraggono"),
- muoversi in direzioni opposte (si "allontanano"),
- oppure essere indipendenti

- Range: Il valore oscilla sempre tra −1 e +1.
- Concordanza (r vicino a +1): Al crescere di X aumenta anche Y. Graficamente, i punti salgono dal basso a sinistra verso l'alto a destra.
- Discordanza (r vicino a −1): Al crescere di X, Y diminuisce. La nuvola di punti scende dall'alto a sinistra verso il basso a destra.
- Assenza di relazione (r vicino a 0): I punti sono sparsi casualmente e non seguono un trend preciso.

Range di oscillazione:
Il coefficiente di correlazione misura proprio questa relazione e può assumere valori tra -1 e +1:

CONCORDANZA r = 1 ➜ correlazione lineare positiva perfetta
Le due variabili sono perfettamente concordi: Al crescere di X aumenta anche Y. Graficamente, i punti salgono dal basso a sinistra verso l'alto a destra.

DISCORDANZA r = -1 ➜ correlazione lineare negativa perfetta
Le variabili sono perfettamente discordi: Al crescere di X, Y diminuisce. La nuvola di punti scende dall'alto a sinistra verso il basso a destra.

r = 0 ➜ assenza di correlazione lineare
Non esiste una relazione lineare tra le due variabili. I punti sono sparsi casualmente e non seguono un trend preciso.

Interpretazione grafica (scatterplot)
Attraverso un diagramma a dispersione (scatterplot) è possibile intuire il tipo di relazione anche prima di calcolare r:
• Se i punti formano una nuvola che va dal basso a sinistra verso l'alto a destra, la correlazione è positiva.
• Se i punti sono quasi allineati ➜ r vicino a 1
• Se la forma è meno precisa ➜$r \approx 0{,}4 - 0{,}5$
• Se i punti vanno dall'alto a sinistra verso il basso a destra, la correlazione è negativa.
• Se sono ben allineati ➜$r \approx -1$
• Se meno definiti ➜$r \approx -0{,}4\ -0{,}5$
• Se i punti sono disposti in modo casuale, senza una forma chiara ➜ r è vicino a 0

EXCEL:

1. **Creare lo scatterplot (grafico a dispersione**
   A. Seleziona l'intervallo dei valori di X e Y
   B. Vai su Inserisci ➜ Grafico
   C. Scegli Grafico a dispersione (XY) oppure grafico a bolle
   D. Seleziona il primo grafico in alto a sinistra

   Obiettivo: capire se esiste una relazione tra calorie assunte e peso.

2. Migliorare la visualizzazione
Per analizzare meglio i dati:
A. Vai su Elementi grafico ➜ Assi ➜ Altre opzioni
B. Si apre il pannello Formato asse
C. Imposta i limiti minimo e massimo
Esempio:
Se le calorie partono da 700, puoi impostare minimo = 800 per osservare meglio la distribuzione.

3. Interpretazione del grafico
Dal grafico possiamo già fare alcune considerazioni:
La correlazione è positiva (i punti salgono da sinistra a destra)
Tuttavia è debole, perché:
• i punti non formano una retta

• la nuvola è abbastanza dispersa
Possiamo stimare:
$r \approx 0{,}4 – 0{,}5$

4. Calcolo manuale della correlazione

Per calcolare il coefficiente di correlazione servono gli z-score.

a. Media

Calcola la somma:
Dividi per il numero di osservazioni N
Esempio:
Totale = 5566
x = 1113{,}2

b. Scarto dalla media

Crea una nuova colonna: xi - media

Simbolo mu **μ**= media della popolazione
Simbolo $\bar{\mathbf{x}}$ = media campionaria

c. Varianza

σ (sigma) = varianza della popolazione
Eleva al quadrato gli scarti: =(scarto)^2
Somma tutti i valori ➜ **devianza**
Dividi per N: Varianza = devianza / N

d. Scarto quadratico medio (deviazione standard)

È la radice della varianza:
=RADQ(varianza)
oppure:
=varianza^(1/2)

Scarto quadratico medio = si ha all'interno della popolazione = qui dividiamo per N
Deviazione standard = si ha all'interno del campione = qui dividiamo per n-1

e. Z-score

Standardizza ogni valore:
z = (x - media) / sqm
Ripeti lo stesso procedimento per Y.

f. Coefficiente di correlazione

Calcola il prodotto:
Zx * Zy
Somma tutti i prodotti:
=SOMMA(intervallo)
Dividi per N:
r = somma / N

5. Metodo veloce in Excel

Puoi evitare tutti i passaggi usando direttamente la funzione:
=CORRELAZIONE(intervallo_X; intervallo_Y)
Dove:
matrice 1 = valori di X
matrice 2 = valori di Y

Esercizio 1

1) Rappresentare graficamente la correlazione tra X ed Y tramite lo scatterplot (diagramma a dispersione)
2) Commentare il grafico e individuare il tipo di correlazione,
3) Stimare il coefficiente di variazione ρ(x,y) senza calcoli
4) Calcolate il Coefficiente di Correlazione Lineare

	X: CALORIE	Y: PESO								
ID	**X**	**Y**	**xi-media**	**(xi-media)^2**	**Z-score x**	**yi-media**	**(yi-media)^2**	**Z-score y**	**Zxi * Z yi**	
1	929	66	-184,2	33929,64	-0,9623	1,20	1,44	0,33859959	-0,32583	
2	1035	61	-78,2	6115,24	-0,40853	-3,80	14,44	-1,072232	0,438043	
3	1366	64	252,8	63907,84	1,320682	-0,80	0,64	-0,2257331	-0,29812	
4	1317	71	203,8	41534,44	1,064696	6,20	38,44	1,74943121	1,862612	
5	919	62	-194,2	37713,64	-1,01454	-2,80	7,84	-0,7900657	0,801556	
Totale	5566	324				0,00			2,478254	
Media	1113,2	64,8								
			Devianza x	183200,8		**Varianza x**	36640,16		**S.q.m. x**	191,4162
			Devianza y	62,80		**Varianza y**	12,56		**S.q.m. Y**	3,544009
COEFF. DI CORRELAZIONE =			0,495651							

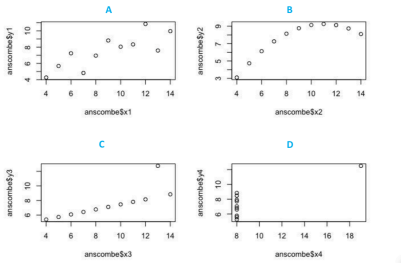
27/04

Ancora sulla correlazione..quando un numero da solo non è sufficiente…

PARADOSSO DI ANSCOMBE

Il Paradosso di Anscombe mostra che, soprattutto nelle scienze sociali ed economiche, è possibile avere dataset molto diversi tra loro che però producono lo stesso coefficiente di correlazione.

Dataset	Correlazione tra	Coefficiente di correlazione
A	x1 e y1	0.816
B	x2 e y2	0.816
C	x3 e y3	0.816
D	x4 e y4	0.816



Questo significa che dati con distribuzioni completamente differenti possono avere lo stesso valore di r suggerendo (in apparenza) una relazione simile tra le variabili.
Tuttavia, osservando la rappresentazione grafica (lo scatterplot), emergono differenze fondamentali.

Perché si parla di "paradosso"?
Si parla di paradosso perché:

- Il coefficiente di correlazione può essere alto (vicino a 1) => suggerisce una forte relazione lineare positiva
- Ma lo scatterplot può mostrare:
  - assenza di relazione
  - relazione non lineare
  - presenza di outlier che falsano il risultato

In altre parole: il numero da solo può essere fuorviante.
Il paradosso insegna una regola fondamentale: il coefficiente di correlazione e lo scatterplot devono essere sempre analizzati insieme

Ordine corretto di analisi:
A. Osservare lo scatterplot
B. Capire se esiste una relazione (e di che tipo)
C. Stimare visivamente l'intensità
D. Calcolare il coefficiente di correlazione

Dalla correlazione alla regressione
Osservando una relazione tra due variabili, può nascere una domanda più profonda: se le due variabili sono correlate, esiste anche un rapporto di causa-effetto?

Ad esempio: *se all'aumentare di X aumenta anche Y, possiamo dire che X causa Y? E se sì, di quanto varia Y al variare di X?*

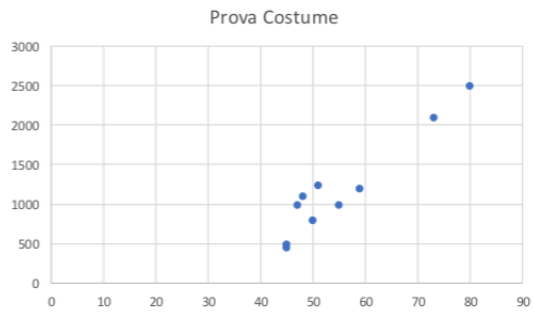
La **REGRESSIONE** risponde proprio a queste domande.
A differenza della correlazione, che misura solo l'intensità e la direzione del legame, la regressione assume un'ipotesi più forte: esiste una relazione di dipendenza (causa-effetto)

In questo contesto:
X = variabile indipendente (causa, antecedente)
Y = variabile dipendente (effetto, conseguente)

! Attenzione: Stabilire quale variabile sia causa e quale effetto non è automatico: dipende dal contesto teorico e dall'interpretazione del fenomeno. Non è il calcolo a dirlo, ma il ragionamento.

Esempio: relazione tra peso e calorie
Nel caso peso–calorie, la rappresentazione corretta prevede:
asse X (ascisse) → calorie (variabile indipendente)
asse Y (ordinate) → peso (variabile dipendente)

Id	Y Peso (KG)	X Calorie giornaliere
1	48	1100
2	51	1250
3	45	500
4	55	1000
5	59	1200
6	73	2100
7	47	987
8	80	2500
9	45	450
10	50	800



$\rho=0,951$

Quindi questo grafico non è corretto! Perchè asse x e asse y

Rispetto al grafico della correlazione, nella regressione compare un elemento in più: una **retta** (spesso tratteggiata)
- Nella correlazione osserviamo solo la nuvola di punti, non c'è alcuna retta, perché non assumiamo una relazione di dipendenza.
- Nella regressione compare una **retta di regressione** che rappresenta una relazione funzionale tra le variabili.

Questa non è una retta qualsiasi, ma una retta di regressione, costruita per descrivere al meglio il legame tra x e y.

Dalla covariazione alla dipendenza
Se osserviamo che le variabili covariano (es. all'aumentare delle calorie aumenta il peso), possiamo ipotizzare qualcosa in più: non solo relazione, ma anche dipendenza.

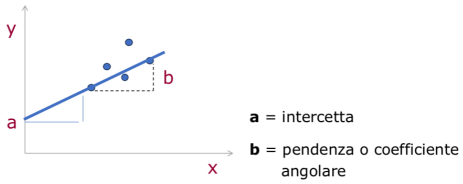
Qui cambia il punto di vista:
Correlazione → relazione "paritaria" tra variabili
Regressione → relazione di causa-effetto

Schema di dipendenza:
A. Modifico X (es. aumento le calorie)
B. Mi aspetto un effetto su Y (variazione del peso)

Quindi:
X = variabile indipendente (causa, predittore)
Y = variabile dipendente (effetto, risposta)

Funzione di regressione
Se scatterplot e correlazione indicano una relazione lineare, possiamo descriverla con una funzione. Questa funzione esprime il legame di dipendenza tra X e Y e permette di prevedere Y conoscendo X. Il coefficiente b rappresenta quanto varia in media Y al variare di una unità di X

Esempio: se b = 0,5 → aumentando le calorie di 1 unità, il peso aumenta in media di 0,5.



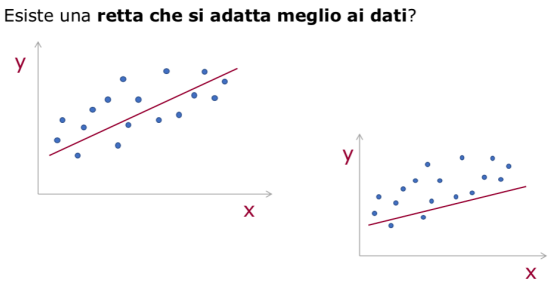
La forma generale di una retta è: $y = a + bx$

Questa equazione ci dice che il valore di y dipende da x:
- il parametro **b** indica di quanto cambia y al variare di x
- mentre **a** rappresenta il valore di y quando x = 0.

In geometria sappiamo che per tracciare una retta bastano due punti. Tuttavia, quando lavoriamo con dati reali rappresentati in uno scatterplot, non possiamo semplicemente scegliere due punti qualsiasi e far passare una retta. I punti, infatti, non sono allineati perfettamente e una retta scelta "a caso" non rappresenterebbe in modo corretto la relazione tra le variabili.

Il problema diventa quindi capire quale sia la retta migliore tra tutte quelle possibili **(retta ai minimi quadrati o di regressione)**

Ci poniamo questa domanda perché, osservando i dati, abbiamo già visto che le due variabili sono correlate, ad esempio con una correlazione positiva.
Questo ci porta a fare un passo in più: non ci interessa più solo sapere che le variabili si muovono insieme, ma vogliamo capire se esiste anche una relazione di dipendenza, cioè se le variazioni di x influenzano quelle di y.



Nel caso ideale, se la relazione fosse perfetta, tutti i punti dello scatterplot si troverebbero esattamente su una retta.
In una situazione di perfetta concordanza positiva, questa retta coinciderebbe con la bisettrice del primo e terzo quadrante: a ogni valore di x corrisponderebbe esattamente lo stesso valore di y.

Tuttavia, nella realtà questo è praticamente impossibile, perché i fenomeni economici e sociali sono sempre influenzati da molti fattori.

Per questo motivo, dobbiamo accontentarci di una soluzione più realistica: cercare una retta che passi il più vicino possibile a tutti i punti. In altre parole, la retta di regressione è quella che minimizza la distanza tra sé e i punti osservati, rappresentando così la migliore approssimazione lineare della relazione tra le variabili.

Tra tutti i punti di uno scatterplot possono passare infinite rette. Il problema è capire quale scegliere per rappresentare al meglio la relazione tra x e y.
La soluzione è utilizzare il **criterio dei minimi quadrati***, che ci permette di individuare la cosiddetta retta di regressione: quella che passa il più vicino possibile a tutti i punti.

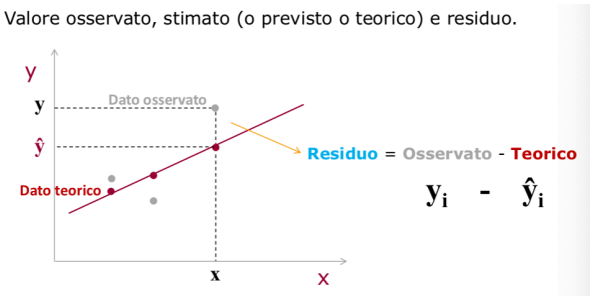
*Def. **criterio dei minimi quadrati**= La retta stimata con il criterio dei minimi quadrati è quella che minimizza la somma dei quadrati degli errori (o residui) tra i valori osservati, y_i, e quelli previsti o stimati, $\hat{y}$

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \text{minimo}$$

In altri termini, è la retta che minimizza i quadrati delle distanze tra tutti i valori osservati (y_i) e quelli teorici ($\hat{y}_i$), in corrispondenza di ciascun valore x_i.

Cosa significa, concretamente, “più vicino possibile”?

Dato osservato e dato teorico
Immaginiamo di avere un punto osservato nello scatterplot, cioè una coppia di valori reali (x, y).
Ora immaginiamo una retta (ancora sconosciuta) che rappresenta la relazione tra le due variabili.
Per ogni valore di x:
- esiste un valore **osservato** y (quello reale)
- esiste un valore **stimato** $\hat{y}$, cioè quello che si troverebbe sulla retta. Questo valore stimato è la proiezione del punto osservato sulla retta.
- La differenza tra valore osservato y e valore stimato $\hat{y}$ si chiama **errore o residuo**.



Se il punto sta esattamente sulla retta => errore = 0
Se il punto è distante => errore > 0
Più grande è questa distanza, meno la retta rappresenta bene quel punto.

L’obiettivo è scegliere la retta che rende questi errori complessivamente più piccoli possibile:
1. si considerano tutte le distanze tra punti osservati e retta
2. si elevano al quadrato (per evitare segni negativi)
3. si sommano

La retta migliore è quella che minimizza la somma dei quadrati degli errori.

La retta di **regressione rappresenta** quindi il miglior compromesso possibile tra tutti i punti e una descrizione “media” della relazione tra x e y.
! È importante capire che nei dati reali i punti non staranno mai perfettamente sulla retta ma ci sarà sempre una certa variabilità. Solo nel caso ideale (quasi mai reale) i valori osservati coinciderebbero perfettamente con quelli teorici. !

Il valore $\hat{y}$ è il valore che ci aspetteremmo se la relazione lineare fosse perfettamente valida.
Ad esempio: per un certo valore di x, la retta può prevedere y = 15 ma il valore reale può essere più alto o più basso. Questo scarto è normale ed è proprio ciò che la regressione cerca di ridurre.

L'equazione della retta di regressione dei minimi quadrati nella **popolazione**:

$$\hat{y} = b_1 x + b_0$$

I parametri della retta
Applicando il metodo dei minimi quadrati otteniamo i parametri della retta:
- b_1 = **coefficiente di regressione** (pendenza) => Indica di quanto varia, in media, y al variare di una unità di x
- b_0 = **intercetta** => Indica il valore di y quando x = 0

I parametri della retta si ottengono attraverso formule che coinvolgono:
- il coefficiente di correlazione r
- le deviazioni standard di x e y
- le medie delle due variabili

In particolare:

Parametro	**Popolazione**	**Campione**
Coefficiente di regressione o **pendenza**	$b_1 = r \cdot \frac{\sigma_y}{\sigma_x}$	$b_1 = r\frac{s_y}{s_x}$
Intercetta	$b_0 = \bar{y} - b_1\bar{x}$	$b_0 = \bar{x}_y - b_1\bar{x}_x$

WorkFlow operativo: Dall'analisi alla regressione
1. **Primo step**: rappresentazione grafica tramite diagramma a dispersione (scatterplot). Questo serve a evitare errori di interpretazione, come quelli messi in evidenza dal Paradosso di Anscombe, dove dati molto diversi possono avere lo stesso coefficiente di correlazione ma relazioni completamente differenti. Osservando il grafico possiamo già capire se esiste una relazione, che tipo di relazione è (positiva, negativa o nulla) e se ha una forma lineare.
2. **Secondo step**: calcolare e interpretare il coefficiente di correlazione lineare r (o ρ). Questo indice misura la direzione e l'intensità della relazione lineare e assume valori compresi tra -1 e +1. Tuttavia, da solo non è sufficiente: ci dice che esiste una relazione, ma non implica necessariamente un legame di dipendenza.
3. **Terzo step**: stima della retta di regressione. Qui si introduce l'idea di dipendenza: si costruisce una funzione che descrive come la variabile Y dipenda da X, utilizzando la retta dei minimi quadrati. Questa retta rappresenta la migliore approssimazione lineare possibile dei dati.
4. **Quarto step:** riguarda il **coefficiente di determinazione R^2***, che misura la bontà di adattamento del modello. In altre parole, risponde a questa domanda: quanto è adeguata la retta nel rappresentare la relazione tra le due variabili?

***Il significato di R^2 (coefficiente di determinazione)**
- si calcola come: $R^2 = r^2$
- Esso varia tra 0 (errore) e 1 (massimo adattamento):

- Se R^2 è vicino a 0, la relazione lineare è debole o non significativa: la retta non è adatta a descrivere i dati.
- Se R^2 è vicino a 1, la retta rappresenta molto bene la relazione: la dipendenza lineare è forte.

In sintesi:
- $R^2 = 0$ => nessun adattamento
- $R^2 = 1$ => adattamento perfetto

Collegamento con l'errore
Il processo di stima della regressione si basa sempre sul confronto tra valore osservato e valore teorico (stimato dalla retta).
- Più questa differenza è grande maggiore è l'errore, peggiore è l'adattamento e più basso sarà R^2.
- Più i valori sono vicini minore è l'errore, migliore è la retta e più alto sarà R^2.

Confronto tra gli indici

Il coefficiente di correlazione **r**: • valori tra -1 e +1 • misura direzione e intensità della relazione	Il coefficiente di determinazione **R^2**: • valori tra 0 e 1 • misura la bontà di adattamento	Il coefficiente angolare b_1: • valori tra $-\infty$ e $+\infty$ • indica l'effetto di X su Y

Tutto il processo si basa su un concetto fondamentale: più la retta è vicina ai dati reali, migliore è il modello e questo si riflette direttamente in un valore alto di R^2.

Se non si trova una retta di regressione significa che le due variabili non rispondono a logiche di linearità oppure potrebbe essere che la variabile indipendente non è la variabile giusta.
La somma delle distanze deve essere la minore possibile. La retta deve minimizzare la distanza tra i punti.

NB. NON TUTTI I FENOMENI POSSONO ESSERE SPIEGATI DALLA DIPENDENZA.

28/04

La regressione
La regressione lineare consente di descrivere e generalizzare la relazione di dipendenza tra due variabili, permettendo di estendere le strutture relazionali osservate nei dati.

Riprendendo l'esempio della profondità e del tempo, è fondamentale analizzare il **residuo**, cioè l'errore di previsione del modello.

Costruzione del residuo
- Si calcolano innanzitutto i **valori stimati (ŷ)** utilizzando l'equazione di regressione.
- Per ogni osservazione si sostituisce il valore di **X osservato** nella retta stimata.
- Il **residuo (e)** è dato da:

$$residuo = valore\ osservato\ (Y) - valore\ stimato\ (\hat{y})$$

Ruolo dei residui
I residui sono fondamentali per verificare l'adeguatezza del modello lineare.
Un modello lineare è considerato appropriato quando:
- I residui sono distribuiti **casualmente attorno allo zero**
- La nube di punti nel grafico residui vs X è **uniforme**, senza pattern evidenti
- La nube di punti forma un ellisse simmetrico

Per dire che il modello lineare è appropriato devo avere una nube di punti uniforme che può essere circoscritta in un ellisse simmetrico, dove i punti sono numericamente simili al di sopra e al di sotto dello zero. I residui sono utilizzati per verificare l'adeguatezza del modello lineare.

R^2 (coefficiente di determinazione)

Il coefficiente di determinazione **R^2** indica la quota di variabilità della variabile dipendente spiegata dal modello.
$R^2 \times 100$ = percentuale di variabilità spiegata
Ad esempio:

- $R^2 = 0,9 \rightarrow$ il modello spiega il 90% della variabilità di Y

⚠ Attenzione: R^2 misura la capacità esplicativa del modello, **non da solo la sua validità**

Ipotesi di omoschedasticità
Perché il modello lineare sia appropriato, è necessario che: la varianza dei residui sia costante per tutti i valori di X
Questa condizione si chiama: omoschedasticità
Significa che:

- ogni osservazione contribuisce con un livello di variabilità simile
- i residui oscillano attorno allo zero in modo **casuale e uniforme**

Eteroschedasticità
Se invece la variabilità dei residui cambia al variare di X (ad esempio aumenta o diminuisce), si parla di: eteroschedasticità
Questo comporta:

- una violazione delle ipotesi della regressione lineare
- una minore affidabilità delle inferenze statistiche (test, intervalli di confidenza)

CORREZIONE IMPORTANTE
“Per non avere varianza la varianza deve essere zero” ✗ FALSO
La varianza NON deve essere zero, ma costante ✔ VERO
Una varianza pari a zero è impossibile nei dati reali e significherebbe assenza totale di errore. Si deve verificare che ogni osservazione porti allo stesso livello di variabilità. Le variazioni nella ricerca della dipendenza dovrebbero oscillare attorno allo zero in modo che ogni unità apporti variabilità costante (senza oscillazioni).

CONCLUSIONE: Un modello lineare è appropriato quando:
✔ i residui sono casuali e centrati attorno allo zero
✔ la loro varianza è costante (omoschedasticità)
✔ non emergono pattern nei grafici dei residui
Al contrario: la presenza di eteroschedasticità indica un modello non adeguato dal punto di vista delle ipotesi statistiche

ESERCIZIO EXCEL:
La formula finale è:
$Y = b1X + b0$
dove:
Y = voto previsto
X = ore di studio
b1 = pendenza della retta
b0 = intercetta, cioè punto di partenza della retta

1. Dati di partenza
   Hai 10 studenti. Per ciascuno hai:
   Ore di studio X Voto Y
   Lo scopo è capire se chi studia più ore prende voti più alti.

2. Totale e media
   Il file calcola prima:
   Totale ore = 309 Totale voti = 162
   Poi divide per 10 studenti:
   Media X = 30,9 Media Y = 16,2
   Quindi, mediamente, gli studenti studiano **30,9 ore** e prendono **16,2**.

3. Scarto dalla media

Poi calcola:

x – media X

y – media Y

Questo serve per capire quanto ogni studente si allontana dalla media.

Esempio studente 1:

Ore = 31 Media ore = 30,9 => 31 – 30,9 = 0,1

Voto = 17 Media voto = 16,2 => 17 – 16,2 = 0,8

Quindi lo studente 1 ha studiato leggermente più della media e ha preso un voto leggermente più alto della media.

4. Scarti al quadrato

Poi il file calcola:

$(x - \text{media } X)^2$

$(y - \text{media } Y)^2$

Perché si elevano al quadrato?

Per due motivi:

1. per eliminare i segni negativi;
2. per misurare quanto i valori sono dispersi rispetto alla media.

5. Devianza

La **devianza** è la somma degli scarti al quadrato.

Nel file:

Devianza X = 736,9

Devianza Y = 177,6

Significa che le ore di studio sono più disperse rispetto ai voti.

6. Varianza

La varianza è: devianza / numero di osservazioni

Nel file:

Varianza X = 73,69

Varianza Y = 17,76

Serve sempre a misurare la variabilità.

7. Scarto quadratico medio

Lo **sqm**, cioè lo scarto quadratico medio, è la radice della varianza.

Nel file:

sqm X = 8,58 sqm Y = 4,21

Significa che, in media, le ore di studio si allontanano dalla media di circa **8,58 ore**, mentre i voti si allontanano dalla media di circa **4,21 punti**.

8. Z-score: zx e zy

Poi il file calcola i valori standardizzati:

zx = (x – media X) / sqm X

zy = (y – media Y) / sqm Y

Questa trasformazione serve per mettere X e Y sulla stessa scala.

Così puoi confrontare ore e voti anche se hanno unità diverse.

9. Prodotto zx × zy

Poi viene calcolato: $zx \times zy$

Questo serve per arrivare alla **correlazione**. Se entrambi sono positivi o entrambi negativi, il prodotto è positivo. Vuol dire che X e Y si muovono nella stessa direzione. Se uno è positivo e l'altro negativo, il prodotto è negativo. Vuol dire che X e Y si muovono in direzioni opposte.

10. Correlazione

Nel file: correlazione = 0,343

La correlazione può andare da **–1 a +1**. In questo caso è positiva, quindi: più ore di studio tendono ad associarsi a voti più alti.

Però 0,343 non è altissimo: la relazione è **positiva ma debole/moderata**.

11. R^2
    Il file poi calcola: R^2 = correlazione^2
    Quindi: $0{,}343^2 = 0{,}118$
    Questo significa che le ore di studio spiegano circa: 11,8% della variabilità del voto
    Quindi il modello spiega poco. Le ore contano, ma probabilmente contano anche altri fattori: metodo di studio, difficoltà dell'esame, ansia, preparazione precedente, qualità dello studio.

12. Coefficiente b1
    Il coefficiente **b1** è la pendenza della retta.
    Nel file:
    b1 = 0,1685
    Significa che per ogni ora in più di studio, il voto previsto aumenta di circa:
    0,17 punti
    Esempio: se uno studia 10 ore in più, il voto previsto aumenta di circa:
    $0{,}17 \times 10 = 1{,}7$ punti

13. Coefficiente b0
    Il coefficiente **b0** è l'intercetta.
    Nel file:
    b0 = 10,992
    Significa che la retta parte da circa 10,99 quando X = 0.
    Quindi la formula diventa:
    Y = 0,1685X + 10,992

14. Y previsto
    La colonna **Y previsto** calcola il voto stimato dalla retta.
    Esempio studente 1:
    X = 31
    Y previsto = $0{,}1685 \times 31 + 10{,}992$
    Y previsto = 16,22
    Il voto reale era 17.

15. Residuo
    Il residuo è:
    Y reale – Y previsto
    Per lo studente 1:
    17 – 16,22 = 0,78
    Quindi ha preso **0,78 punti in più** rispetto a quanto previsto dal modello. Se il residuo è positivo: lo studente ha fatto meglio del previsto. Se il residuo è negativo: ha fatto peggio del previsto.

16. Grafici dei residui
    In basso il file prepara due controlli:
    X / residuo
    Y / residuo
    Servono per vedere se il modello lineare è adeguato. Idealmente i residui dovrebbero distribuirsi in modo casuale intorno allo zero. Se invece formano un andamento preciso, vuol dire che la retta non descrive bene i dati.

Conclusione finale
La retta trovata è:
Y = 0,1685X + 10,992
Quindi esiste una relazione positiva tra ore di studio e voto: chi studia di più tende a prendere un voto più alto. Però il legame è debole, perché: $R^2 = 0{,}118$
cioè le ore di studio spiegano solo l'**11,8%** delle differenze nei voti. (Eteroschedasticità)

4/05

(Ripasso) Statistica bivariata
Relazione tra due fenomeni qualitativi*=/ quantitativi
=> in questo caso non possiamo utilizzare correlazione e regressione e stimare il legame e la forza di questo legame

- **correlazione** lineare semplice: indagare l'interdipendenza di due variabili x e y => co-variare di due variabili contemporaneamente studiandone il comportamento
- **regressione** lineare: si fa un passo successivo studiando la dipendenza tra 2 variabili. In cui x spiega ogni cambiamento unitario di y => rapporto causa-effetto). Attraverso l'equazione della retta di regressione è possibile fare previsioni sostituendo il valore di x per trovare y.

Lo step visivo si basa sulla rappresentazione grafica dello scatterplot, che permette di capire se esiste una relazione lineare, positiva o negativa, e, in base alla dispersione dei puntini, se la relazione è forte o debole.
Il coefficiente di correlazione stima la forza della relazione (-1 +1) lineare di Pearson. Se il coefficiente di Pearson è pari a zero, vuol dire che siamo in una situazione di assenza di relazione lineare.

Introduzione
Esempio 1: chi acquista birra, acquista anche sacchetti di patatine?
Esempio 2: chi acquista latte scremato, acquista anche merendine al cioccolato?
Esempio 3: gli abitanti di Lazio, Lombardia, Liguria e Piemonte tifano solo rispettivamente per Lazio, Inter, Samp e Torino?
Esempio 4: chi ha un reddito più elevato è più felice rispetto a chi ha un reddito basso?

Tutte queste variabili hanno delle modalità che non sono quantitative (numeriche) ma **qualitative** => come stimare la forza di 2 variabili su scala ordinale o nominale?
Dobbiamo costruire una configurazione di tabella che riesca a incrociare i valori di due variabili
Tabella di contingenza o doppia entrata

Esempio 1 - Tavola di contingenza sesso/abitudine di guardare la tv
Mette in relazione contemporaneamente:
- variabile genere nelle modalità maschio / femmina
- variabile abitudine di guardare la tv la sera nelle modalità no / sì

Conteggio

x		abitudine di guardare la tv		Totale
		no	si	
sesso	maschio	11	459	470
	femmina	15	506	521
Totale		26	965	991

Quindi stiamo osservando l'abitudine di guardare la tv la sera a condizione di essere un maschio o una femmina.

La tabella di contingenza va ad un livello più approfondito, in questo caso va ad incrociare l'abitudine di guardare la tv prima di andare a letto differenziando se il soggetto è di genere femminile o maschile.

A	B	C	D	E	F
Reddito \ livello di felicità	basso	medio	alto	totale	
basso	21	159	110	290	
medio	53	372	221	646	
alto	94	249	83	426	
totale	168	780	414	1362	N o collettivo
Reddito \ livello di felicità	basso	medio	alto	totale	
basso	n11	n12	n13	n1.	
medio	n21	n22	n23	n2.	
alto	n31	n32	n33	n3.	
totale	n.1	n.2	n.3	N	

11, 459, 15, 506 sono frequenze[4] del col lettivo[5] che presentano congiuntamente la modalità i-esima di un carattere e la modalità i-esima di un secondo carattere rilevate entrambe sulla stessa unità statistica o componente. Si parla di **FREQUENZE CONGIUNTE,** che vanno a vedere l'essere maschio o femmina e al contempo guardare o non guardare la tv.
Notazione scientifica
- frequenze assolute univariate: ni (frequenza relative alla variabile i)
- **Frequenze congiunte**: nij[6]
  i= variabile sesso (modalità maschio / femmina)
  j= modalità no/sì

Distribuzioni marginali
La tabella, oltre alle celle congiunte, ci fornisce un'altro aspetto informativo dai totali di riga e dal totale di colonna in gergo totali marginali, perchè al margine della tabella
- Totale di riga: uomini e donne intervistate
- Totale di colonna: chi ha risposto no (26) e sì ()

Il collettivo (N) è la somma delle distribuzioni marginali.

Quindi
- 11 e 15 sono una **distribuzione condizionata**: abitudine di non guardare la tv a condizione di essere maschio
- 459 e 506 sono anch'essi una distribuzione condizionata: abitudine di non guardare la tv a condizione di essere femmina

Esempio 2
Mette in evidenza il legame tra 2 variabili (reddito e livello di felicità)

Reddito	**Livello di felicità**			
	BASSO	MEDIO	ALTO	Tot.
BASSO	21	159	110	290
MEDIO	53	372	221	646
ALTO	94	249	83	426
Tot.	168	780	414	1362

Dalla tabella di contingenza posso andare ad utilizzare solo i totali di riga e colonna (distribuzioni marginali) e costruire altre 2 tabelle

Totale di riga = se voglio individuare soltanto le persone che hanno i diversi livelli di reddito (carattere x

Reddito		
Basso	290	n1.
Medio	646	n2.
Alto	426	n3.
Totale	1362	N

Totale di colonna = se voglio individuare soltanto le persone che hanno diversi livelli di felicità (carattere y)

Liv. Di felicità		
Basso	168	n.1
Medio	780	n.2
Alto	4141	n.3
Totale	1362	N

[4] Modalità: guardare la tv a condizione di essere maschio/femmina
Frequenza: quante volte si frequenta una modalità

[5] il collettivo si riferisce a campione/popolazione

[6] si legge n con ij

Notazione scientifica, il “.” sostituisce la modalità ignota a seconda che sia x o y
- Totale di riga **n1.** Perchè deriva da una frequenza congiunta ma rimane ignota la modalità di y
- Totale di colonna **n.1** deriva da una frequenza congiunta ma rimane ignota la modalità di y

Rappresentazione grafica

Reddito \ livello di felicità	basso	medio	alto	totale	fij=nij/N
basso	0,02	0,12	0,08	0,21	
medio	0,04	0,27	0,16	0,47	
alto	0,07	0,18	0,06	0,31	
totale	0,12	0,57	0,30	1,00	

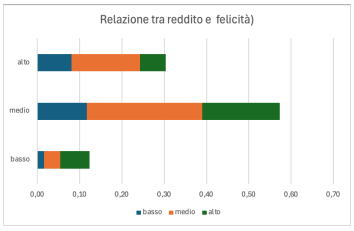


Grafico a barre in pila[7]

Per una correttezza del grafico si utilizzano frequenze relative o percentuali e non assolute. Anche per le frequenze congiunte posso calcolare le frequenza relative, tramite la formula:

$$f_{ij} = \frac{n_{ij}}{N}$$

Analisi dell'associazione per caratteri qualitativi
Studio delle relazioni tra caratteri
- Indipendenza:
  - Due caratteri sono statisticamente indipendenti quando la conoscenza delle modalità di uno non consente di prevedere le modalità dell’altro
- Dipendenza
  - Due caratteri sono dipendenti quando si può stabilire un legame tra le modalità di un carattere e quelle di un altro
- Massima dipendenza :
  - Due caratteri sono interdipendenti quando si può stabilire un legame bidirezionale tra le modalità di un carattere e quelle di un altro. (cioè, ad ogni modalità del carattere X corrisponde una sola modalità del carattere Y e viceversa)

Esempio 2: c’è associazione tra livello di istruzione e genere?
I maschi con livello di istruzione alto sono più delle femmine. Ma le numerosità dei maschi e femmine sono diverse!

Livello Istruzione	Genere		Totale
	Maschi	Femmine	
basso	300	100	400
alto	300	200	500
Totale	600	300	900

Livello Istruzione	Genere	
	Maschi	Femmine
basso	50	33.33
alto	50	66.67
Totale	100	100

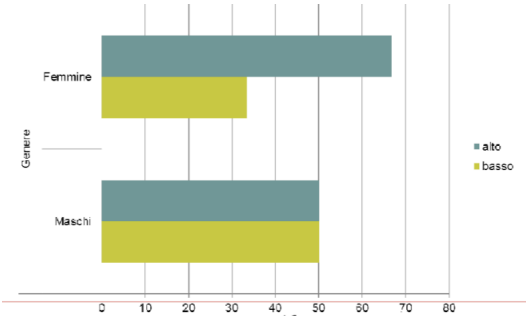
E' opportuno calcolare le frequenze relative del livello di istruzione condizionate al genere. In questo modo abbiamo omogeneizzato i totali.

Si vede che tra le donne l a percentuale di coloro che hanno un livello di istruzione elevato è più alta. Le distribuzioni condizionate sono diverse. Si parla di associazione! Se so che una persona è donna è più probabile che il suo livello di istruzione sia alto rispetto al caso in ci si riferisca ad un maschio.

[7] chiamato erroneamente da excel come “Istogramma”, utilizzato x dati quantitativi organizzarti in classi

Per rappresentare graficamente l'associazione tra variabili qualitative, utilizzerò il grafico a barre percentuali impilate (a pila)

Perché è il grafico corretto per variabili qualitative?
Perché le variabili qualitative non hanno media o andamento continuo.



Non possiamo usare:
- istogrammi continui
- scatterplot
- rette di regressione

Dobbiamo confrontare:
•categorie
- proporzioni
- distribuzioni interne

E le frequenze percentuali condizionate servono esattamente a questo.

1. Distribuzione condizionata di X rispetto a Y
   Vuol dire: "Fissata una modalità di Y, come si distribuisce X?"
2. Distribuzione condizionata di Y rispetto a X
   Qui facciamo il contrario.Vuol dire: "Fissata una modalità di X, come si distribuisce Y?"

La parola "**condizionata**" significa: "**considerando solo un sottogruppo**"
Infatti:
- se condiziono rispetto a Y → scelgo una colonna
- se condiziono rispetto a X → scelgo una riga

! Differenza con distribuzione marginale !
La distribuzione marginale usa TUTTI i dati senza condizioni. Nella distribuzione condizionata invece restringi il campione a un gruppo specifico.

05/05

Relazioni tra variabili qualitative
RIPETIZIONE: differenza tra relazione indipendente, dipendente e dipendente al massimo

Per calcolare l'associazione delle variabili qualitative dobbiamo seguire tre step:
1. Frequenze teoriche
2. Contingenze
3. Indice Chi quadro normalizzato: calcola la connessione tra variabili qualitative

1. Frequenze teoriche: come si costruiscono?
- Prendo in riferimento una frequenza assoluta
- La formula per calcolare la rispettiva frequenza teorica è:

totale della colonna corrispondente * totale della riga corrispondente
/
totale delle osservazioni (ultimo numero in fondo a destra che indica il totale complessivo di tutte le celle)

$$\hat{n}_{ij} = \frac{n_{.j} \times n_{i.}}{N}$$

Si parla di Indipendenza stocastica nel caso in cui le frequenze osservate sono uguali alle frequenze teoriche

Confronto le frequenze osservate con le frequenze teoriche avendo cura di rispettare le posizioni. Se i due caratteri osservati non sono uguali, non c'è indipendenza stocastica.

$$n_{ij} = \hat{n}_{ij}$$

Tabella di frequenze empiriche assolute

Regioni	Lazio	Inter	Samp	Torino	Totale
Lazio	23	8	5	4	40
Lombardia	34	13	15	3	65
Liguria	44	2	1	14	61
Piemonte	5	6	33	24	68
Totale	106	29	54	45	234

Esempio:
- Se si fa riferimento alla frequenza congiunta n_{32}

Ottengo la frequenza teorica facendo: 29 (totale colonna) * 61 (totale riga) / 234 = 7,56

NB: per verificare la posizione corretta delle frequenze contingenti il PRIMO numero si riferisce sempre alla posizione della variabile in riga e il secondo alla posizione della variabile in colonna

2. Contingenze:
La differenza tra frequenze congiunte osservate e frequenze congiunte teoriche si chiama CONTINGENZA e si indica con c_{ij}
Il fenomeno della contingenza è equivalente al residuo o errore calcolato in base alla differenza tra ŷ teorico e y osservato)

$$c_{ij} = n_{ij} - \hat{n}_{ij}$$

La somma delle contingenze così come la somma dei residui è pari a zero.

3. Indice di contingenza x^2
Come si misura l'associazione di caratteri qualitativi?
Attraverso il calcolo dell'indice del chi - quadro e l'indice del chi - quadro normalizzato
La misura dell'intensità della connessione tra i caratteri è data dall'indice di contingenza di Pearson (detto anche indie del chi - quadro)
x^2 = sommatoria delle righe * sommatoria delle colonne * c^2 / n^ teorico (freq. teoriche)

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(c_{ij}^2)}{\hat{n}_{ij}}$$

Cosa mi serve per calcolare l'indice di contingenza?

1. È necessario creare la **tabella delle contingenze al quadrato** e considerare affianco la **tabella delle frequenze teoriche**
2. Creo così un'unica tabella del **RAPPORTO TRA CONTINGENZE AL QUADRATO / FREQUENZE TEORICHE**
3. Dopo aver completato la tabella finale, il mio indice *chi quadro* x^2 sarà la SOMMA delle celle di questa tabella → *funzione excel =SOMMA (intervallo delle celle*)

A questo punto devo capire il grado di connessione ricavabile dall'indice ottenuto:
Per farlo, ho bisogno di paragonarlo al valore massimo che può assumere l'indice x^2, che si può ottenere con un'operazione standard:

MAX = N * k
Legenda:
k = valore minimo tra il numero delle righe e il numero delle colonne della tabella a doppia entrata
meno 1
k = MIN (r-1, c-1)

Esempio: se in una tabella qualsiasi ho 5 righe e 4 colonne, considero 4 come numero minore e sottraggo 1

Indice normalizzato di contingenza

$$\tilde{\chi}^2 = \frac{\chi^2}{MAX} \qquad 0 \le \tilde{\chi}^2 \le 1$$

Se x^2 normalizzato è vicino a zero = connessione debole
Se x^2 normalizzato è vicino a 1 = connessione forte

Continuando sull'esempio di prima:
- k = min (4-1 , 4-1) = (3,3) = 3
- MAX = 234*3 =702

L'indice del x^2 (chi quadro) normalizzato = 94,75/ 702 = 0,135 → è vicino allo zero quindi il legame tra squadra che tifo e regione in cui abito è **debole**

EXCEL:

Frequenze teoriche				
	Lazio	Inter	Samp	Torino
Lazio	18.12	4.957	9.231	7.692
Lombardia	29.44	8.056	15	12.5
Liguria	27.63	7.56	14.08	11.73
Piemonte	30.8	8.427	15.69	13.08

Contingenze=frequenza empiriche-teoriche				
	Lazio	Inter	Samp	Torino
Lazio	4.88	3.043	-4.231	-3.692
Lombardia	4.556	4.944	0	-9.5
Liguria	16.37	-5.56	-13.08	2.269
Piemonte	-25.8	-2.427	17.31	10.92

Contingenze al quadrato				
	Lazio	Inter	Samp	Torino
Lazio	23.82	9.258	17.9	13.63
Lombardia	20.75	24.45	0	90.25
Liguria	267.9	30.91	171	5.149
Piemonte	665.8	5.892	299.6	119.3

Rapporto tra contingenze al quadrato e frequenze teoriche					
	Lazio	Inter	Samp	Torino	totale
Lazio	1.314	1.868	1.939	1.772	6.893
Lombardia	0.705	3.035	0	7.22	10.96
Liguria	9.695	4.089	12.15	0.439	26.37
Piemonte	21.62	0.699	19.09	9.124	50.53
totale	33.33	9.691	33.18	18.56	94.75

=$F2*B$6/F6

Formula per bloccare le frequenze teoriche (unici valori da dover bloccare) per non ripetere ogni volta la formula:
Il simbolo $ permette di scegliere quali riferimenti devono rimanere fissi durante il trascinamento della formula:

- **$F2**
  La colonna F è bloccata, mentre il numero della riga può cambiare.
  Questo significa che, spostandosi verso destra, Excel continuerà sempre a prendere i valori della colonna F (i totali di riga).
- **B$6**
  In questo caso è bloccata la riga 6, mentre la colonna può cambiare.
  Quindi, trascinando la formula lateralmente, Excel continuerà a usare sempre i valori della riga dei totali di colonna.
- **F6**
  Qui sono bloccati sia la colonna sia la riga.
  Questo riferimento resta sempre identico perché rappresenta il totale generale della tabella, che non deve mai cambiare.

CAPITOLO 3: SINTETIZZARE NUMERICAMENTE I DATI

16/03

MEDIANA E QUARTILI - Misure di sintesi
Una misura di sintesi tradizionalmente utilizzata è la media, che però presenta un limite:
non si traduce sempre in modo veritiero perché risente di boost, valori molto alti che modificano l'andamento regolare.

Quando si presentano modalità distanti e lontane dai valori costanti vengono dette anomale o outlier.

La mediana è definita media lasca e non risente dei valori anomali.

La famiglia delle misure di sintesi basate sulle posizioni è composta da:
- Primo quartile 25% Q1
- Secondo quartile o mediana (50%) Q2
- Terzo quartile 75% Q3

Se siamo interessati a descrivere il comportamento di parti del campione utilizziamo questi strumenti (es. quando lavoriamo in termini di percentuali; il primo 25% ha questo atteggiamento, la prima metà 50% ha questo atteggiamento, ecc.)

La mediana è una MISURA ROBUSTA perché NON è influenzata dai valori estremi mentre. la media NON è robusta perché risente dei valori outlier

=> Non si può sovrapporre il concetto di media e mediana perché riportano a misure differenti.

Un valore sintetico descrive parzialmente la variabile:
Media aritmetica: è la misura più comune e si calcola solo su dati quantitativi.
Consiste nel: sommare tutti i valori e dividere per il numero di osservazioni.
- Se calcolata sulla popolazione —> si chiama parametro (μ)
- Se calcolata su un campione —> si chiama statistica ($\bar{x}$)

La **moda** è il valore che compare più frequentemente. Si può calcolare sia per dati quantitativi che qualitativi. È semplicemente la categoria o valore più comune.

- Mediana e quartili: variabili quantitative e ordinali

Il concetto di frequenza cumulata sottende che le frequenze coincidano con il concetto di posizione: le prime 100 tacche corrispondono ad un grado e corrispondono alle prime 100 posizioni.
→ La frequenza si può vedere come posizione.

La posizione si può esprimere in termini percentuali: alla posizione che si trova al 50% corrisponde un determinato valore - modalità (es. un valore per esprimere l'altezza).
La mediana divide in due parti uguali il mio collettivo e mi permette di far coincidere la posizione con la frequenza cumulata (che è serie ordinata progressiva).

La mediana di una variabile è il valore che occupa la posizione centrale e si può trovare sia nelle numerosità pari che dispari. Nelle numerosità pari per determinare la modalità che corrisponde alla posizione mediana avverrà il fenomeno di interpolazione.

I percentili dividono una distribuzione ordinata in 100 parti uguali. Si lavora sulle posizioni delle modalità che corrispondo alle frequenze.
Es. il 25% degli studenti ha avuto 30 all'esame di statistica

- 25esimo percentile corrisponde a 25/100 = 1/4
- 50esimo corrisponde a 50/100 = 1/2
- 75esimo corrisponde a 75/100 = 3/4

$i = 1 + (n-1) * p$ (percentile) calcolo della posizione del campione
$i = 1 + (N-1) * p$ calcolo della posizione della popolazione

N maiuscolo o minuscolo va a indicare il totale delle frequenze, perciò valgono entrambe le formule, è sufficiente prendere il totale (non cambia)

Variabile X: voto Statistica

x_i	n_i	Ni
19	2	2
20	1	3
21	1	4
22	1	5
23	1	6
25	1	7
Totale	7	

i=1+(7-1)0,5=
i=1+(6*0,5)=
i=1+3= **4**

Voglio costruire la mediana che corrisponde al percentile 0,5

I quartili hanno dei percentili definiti:
- Q1 = 0,25
- Q2 = 0,5
- Q3 = 0,75

Faccio → i = 1 + (7-1) 0,5 = 4 → perciò prendo in considerazione la posizione 4 e osservo a quale variabile x1 corrisponde

In questo caso la mediana è 21 → 21/30 è il voto che crea due sottogruppi di uguale numerosità
Devo sempre creare la colonna delle frequenze cumulate

Se voglio calcolare Q1, faccio → i = 1 (7-1) 0,25 = 2,5
La posizione 2,5 ricade tra 2 e 3, ma seguendo il metodo dell'interpolazione:
Il 2,5 si compone di una parte intera 2 e una parte decimale 0,5 e si trova tra due modalità che sono 19 (x1) e 20 (x2)
Q1 = x1 + 0,5 (x2 - x1) → formula computazionale generalizzata iesima
Q1 = 19 + 0,5 (20-19) = 19 + 0,5 (1) = 19,5

Dopo aver descritto il valore "tipico" dei dati (con media, mediana, moda), bisogna capire quanto i dati sono dispersi, cioè quanto si allontanano dalla media.
Le principali misure di dispersione sono:
1. Intervallo di variazione
2. Varianza
3. Deviazione standard

Intervallo di variazione (range): è la misura più semplice. Si calcola facendo = valore massimo - valore minimo.
Quindi considera solo i due estremi della distribuzione.
Limiti importanti:
- Dipende molto dai valori estremi
- Usa solo due dati → quindi è poco rappresentativo

Per questo si dice che non è robusto.

Varianza: misura la dispersione considerando tutti i dati. L'idea di base è vedere quanto ogni valore si allontana dalla media
Procedimento concettuale:
1. Si calcola la distanza di ogni valore dalla media
2. Si eleva al quadrato (per evitare segni negativi)
3. Si fa la media di questi valori

Interpretazione:
- Più i dati sono lontani dalla media → più la varianza è alta
- Dati concentrati → varianza bassa

Varianza: popolazione vs campione
Come per la media, bisogna distinguere:
- Varianza della popolazione (σ^2) → si divide per N
- Varianza del campione (s^2) → si divide per n – 1

Perché n - 1?
- Se dividessimo per n, la varianza sarebbe sottostimata
- Usare n - 1 serve a correggere questa distorsione
- Si parla di stimatore non distorto

n - 1 = gradi di libertà
(perché l'ultimo valore è vincolato dagli altri e dalla media)

Deviazione standard: è la radice quadrata della varianza. Rappresenta la distanza tipica dei dati dalla media.
Serve per risolvere il problema della varianza:
- Torna nelle unità originali dei dati
- Quindi è più interpretabile

Più è grande: più i dati sono dispersi
Più è piccola: più i dati sono concentrati intorno alla media
→ Si possono confrontare deviazioni standard solo se le unità di misura sono le stesse

Distribuzione campanulare (idea finale): se i dati hanno forma simmetrica a campana, la deviazione standard diventa ancora più utile. Permette di capire quanti dati stanno vicino alla media, usando intervalli del tipo: **media ± k deviazioni standard**.
Serve quindi per descrivere la distribuzione in modo più preciso.
In sintesi:
- Range → semplice ma poco affidabile
- Varianza → precisa ma poco intuitiva
- Deviazione standard —> la più utile nella pratical percintili sono un altro modo di analizzare la popolazione statistica che dividono la distribuzione ordinata in 100 parti uguali.

Le **misure di posizione**, servono a capire dove si colloca un valore rispetto agli altri dati. Non descrivono quindi "quanto vale" o "quanto varia", ma la posizione relativa nella distribuzione.

1) Z-score (punteggi standardizzati): serve per confrontare valori anche appartenenti a distribuzioni diverse. Indica di quante deviazioni standard un valore si trova sopra o sotto la media. Serve per confronti corretti, anche tra contesti diversi.
Caratteristiche:
- Non ha unità di misura
- Ha media = 0 e deviazione standard = 1

Interpretazione:
- $z > 0$ → valore sopra la media
- $z < 0$ → valore sotto la media
- $z = 0$ → valore uguale alla media

Esempio:
- $z = 1.24$ → valore sopra la media di 1.24 deviazioni standard
- $z = –2.31$ → valore sotto la media di 2.31 deviazioni standard

Problema: la distribuzione di frequenza della Tabella 3.10 riporta i dati di un campione casuale di 40 tassi di interesse a 3 anni relativi a fondi comuni d'investimento. Approssimare la media dei tassi di interesse a 3 anni.
Metodo: per trovare il valore approssimato della media seguiamo i seguenti step.

Step 1: calcoliamo il valore centrale di ogni classe. Il valore centrale si trova sommando gli estremi inferiori di due classi consecutive e dividendo il risultato per 2.
Step 2: moltiplichiamo il valore centrale per la frequenza per ottenere $x_i f_i$ per ciascuna classe.
Step 3: calcoliamo $\Sigma x_i f_i$.
Step 4: calcoliamo la somma delle frequenze, Σf_i.
Step 5: inseriamo i valori ottenuti nella Formula (8) per ottenere la media di una distribuzione di frequenza.
Soluzione
Step 1: l'estremo inferiore della prima classe è 10. L'estremo inferiore della seconda classe è 12. Quindi il valore centrale della prima classe è (10 + 12)/2 = 11, quindi x_1 = 11. Gli altri valori centrali sono elencati nella colonna 2 della Tabella 3.11.
Step 2: calcoliamo $x_i f_i$ moltiplicando ciascun valore centrale per la corrispondente frequenza. I risultati sono riportati nella colonna 4 della Tabella 3.11.
Step 3: sommiamo i valori della colonna 4 della Tabella 3.11 per ottenere $\Sigma x_i f_i = 592$.
Step 4: sommiamo le frequenze della colonna 3 per ottenere $\Sigma f_i = 6 + 14 + \ldots + 3 = 40$.

$$\bar{x} = \frac{\Sigma x_i f_i}{\Sigma f_i} = \frac{592}{40} = 14.8$$

Step 5: sostituiamo i valori ottenuti nella Formula (8), per ottenere:

Tabella 3.11

Classe (tasso di interesse a 3 anni)	Valore centrale della classe, x_i	Frequenza, f_i	$x_i f_i$
10-11.99	(10 + 12)/2= 11	6	(11)(6) = 66
12-13.99	13	14	(13) (14) = 182
14-15.99	15	10	150
16-17.99	17	4	68
18-19.99	19	3	57
20-21.99	21	0	0
22-23.99	23	3	69
$\Sigma f_i = 40$	$\Sigma x_i f_i = 592$		

CONTROLLARE LA PRESENZA DI OUTLIER IN UNA DISTRIBUZIONE

Ogni volta che si conducono analisi statistiche è necessario controllare la presenza o meno di valori molto distanti da quelli che caratterizzano la distribuzione. I valori estremi sono anche noti come **outlier**.
Ogni volta che si trovano outlier è necessario investigarne le origini.
Essi si possono presentare casualmente oppure possono essere il risultato di errori durante la misurazione della variabile o in fase di immissione o di campionamento dei dati.

Gli outlier influenzano la media e la deviazione standard, perché nessuna di queste due misure è robusta. Poiché queste misure sono utilizzate molto frequentemente nella statistica inferenziale, ogni conclusione basata su una distribuzione che presenti outlier potrebbe essere inattendibile.

Per controllare la presenza di outlier, utilizziamo i quartili seguendo questi step.
- **Step 1** : determinare il primo e il terzo quartile della distribuzione.
- **Step 2** : calcolare l'intervallo interquartile.
- **Step 3** : determinare i limiti. I **limiti** servono per determinare gli outlier.
  Limite inferiore = $LF = Q_1 - 1.5\ (IQR)$
  Limite superiore = $UF = Q_3 + 1.5\ (IQR)$
- **Step 4** : se una osservazione è più piccola del limite inferiore o più grande del limite superiore, allora viene considerata un outlier.

DIFFERENZA TRA CORRELAZIONE E CASUALITÀ

Correlazione ≠ causalità: una forte relazione tra due variabili non implica che una causi l'altra, soprattutto nei dati osservazionali.
Una **variabile nascosta** è un fattore non considerato che influenza entrambe le variabili, creando una correlazione apparente.
Esempio: la temperatura è una variabile nascosta perché, aumentando, fa crescere sia l'uso dell'aria condizionata sia alcuni reati → le due variabili risultano correlate, ma senza rapporto causale diretto.

MEDIA E MEDIANA NELLE DISTRIBUZIONI IN CLASSI

Quando i dati sono raggruppati in **classi** (es. 0–10, 10–20, 20–30...), **non conosciamo i valori singoli**, ma solo quante osservazioni cadono in ogni intervallo.
Per questo motivo, media e mediana non possono essere calcolate esattamente, ma solo approssimate.

MEDIA PER DISTRIBUZIONI IN CLASSI

Dato che non abbiamo i valori precisi, si fa un'ipotesi: ogni classe viene rappresentata dal suo **valore centrale** (cioè il punto medio dell'intervallo).
Come si trova il valore centrale?
Si calcola così:

$$x_i = \frac{\text{limite inferiore} + \text{limite superiore}}{2}$$

Questo valore rappresenta "in media" tutti i dati di quella classe.
Procedimento logico

1. Per ogni classe calcoli il **valore centrale** x_i
2. Lo moltiplichi per la **frequenza** f_i
   → $x_i \cdot f_i$
3. Sommi tutti i prodotti
4. Dividi per il totale delle frequenze

Formula della media

$$\mu = \frac{\sum x_i f_i}{\sum f_i}$$

Interpretazione importante

- $x_1 f_1$ ≈ somma dei dati della **prima classe**
- $x_2 f_2$ ≈ somma dei dati della **seconda classe**
- ...

Quindi la somma totale: è una **stima della somma reale dei dati**

MEDIANA PER DISTRIBUZIONI IN CLASSI

La **mediana** è il valore che divide i dati in due parti uguali.
Ma con dati in classi non sappiamo il valore preciso e sappiamo solo **in quale classe cade**, quindi dobbiamo **stimarla**

Procedimento logico

1. Calcoli $\frac{N}{2}$
   → posizione della mediana
2. Usi le **frequenze cumulate** per trovare: la **classe mediana** (cioè dove cade la mediana)
3. Poi **interpoli** dentro quella classe

Formula della mediana

$$M = L + \frac{\frac{N}{2} - CF}{f} \cdot i$$

- **L** = limite inferiore della classe mediana
- **N** = numero totale di osservazioni
- **CF** = frequenza cumulata prima della classe mediana
- **f** = frequenza della classe mediana
- **i** = ampiezza della classe

ESEMPIO

Problema: data la distribuzione in classi dei metri quadrati di 500 appartamenti riportata nella successiva tabella, si calcoli la **mediana**.
Step 1: calcolare le frequenze cumulate, indicate con N_i.
Step 2: calcolare dove ricade la posizione mediana; 500 è pari; quindi, la posizione mediana ricadrà tra 250 e 251.
Step 3: individuare L, ossia l'estremo inferiore della classe in cui ricade la mediana, 2000.
Step 4: la numerosità è pari a 500.
Step 5: individuare la frequenza cumulata antecedente alla classe mediana, 179 in questo caso.
Step 6: determinare l'ampiezza della classe mediana, in questo caso 2499 – 2000 = 499.
Step 7: il risultato è 2297; sono i metri quadrati che individuano i due gruppi di abitazioni: quelli di ampiezza contenuta e quelli di grande ampiezza.

Classi di mq	n_i	N_i	
0-499	5	5	
500-999	17	22	
1000-1499	36	58	
1500-1999	121	179	
2000-2499	119	298	Classe mediana
2500-2999	81	379	
3000-3499	47	426	
3500-3999	45	471	
4000-4499	22	493	
4500-4999	7	500	
Totale	**500**		